DOCUMENT RESUME

ED 352 396 TM 019 301

AUTHOR Madaus, George F.; And Others
TITLE Testing and Evaluation: Learning from the Projects We Fund.
INSTITUTION Council for Aid to Education, New York, NY.
SPONS AGENCY Aetna Life and Casualty Foundation, Inc., Hartford, CT.
PUB DATE 92
NOTE 63p.; Paper in the Policy Issues in the Conduct of Corporate Support for Education Series.
AVAILABLE FROM Council for Aid to Education, Publications Department, 51 Madison Avenue, Suite 2200, New York, NY 10010 ($15.00; free only for Council supporters).
PUB TYPE Guides - Non-Classroom Use (055) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Corporate Support; Educational Finance; Endowment Funds; *Evaluation Methods; Evaluation Utilization; *Financial Support; Foundation Programs; Grants; Planning; *Program Evaluation; *Research Projects; *Test Use
IDENTIFIERS Policymakers

ABSTRACT

An overview of issues in program evaluation is provided for corporate contributions administrators, including the place of testing in such evaluations, so that the administrators are in a better position to plan, commission, monitor, evaluate, and use evaluations of the projects they fund. In the first chapter, metaphors and models for educational evaluation are explored, examining how metaphors used to describe the process of education and schools can influence approaches to program evaluation. The three main approaches to program evaluation are described as: (1) goals-oriented; (2) decision-oriented; and (3) responsive or naturalistic. The second chapter offers practical advice on commissioning, monitoring, evaluating, and using evaluations. The third chapter considers educational tests and issues related to their use in program evaluation. Seven figures and five tables illustrate the discussion. A glossary is included, and three excerpts of the report in card form are enclosed. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

COUNCIL FOR AID TO EDUCATION

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

PRISCILLA LEWIS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

Policy Issues in the Conduct of Corporate Support for Education

Testing and Evaluation

Learning from the Projects We Fund

George F. Madaus
Walt Haney
Amelia Kreitzer

TM019301

BEST COPY AVAILABLE

About the Council

The Council for Aid to Education, Inc., is a tax-exempt, not-for-profit corporation dedicated to promoting an effective partnership between corporate and foundation funders and the education community in the task of improving America's educational system.

Founded in 1952 by the chairmen of Standard Oil Company of New Jersey (Exxon), General Motors, U.S. Steel (USX Corporation), the Container Corporation of America, and the Armstrong Cork Company, the Council was originally designed to encourage private-sector support of higher education. Today the Council has expanded its mission to become a resource for companies and private foundations interested in making effective contributions to the quality of American precollege and higher education. This broadened mission also includes the provision of a range of information services to educational institutions interested in cultivating private support.

Supported by large and small corporations as well as by private foundations, the Council helps education funders target key areas of educational concern and identify promising approaches to educational improvement. It points corporations and foundations toward the issues and opportunities in which they have a compelling interest and on which they can have an impact, and alerts funders to the implications of emerging educational challenges. By tracking trends and themes in private support of education over time, the Council can help private funders and educational institutions communicate with each other more clearly, plan more effectively, and assess the results of their efforts in a larger and more meaningful context. Through its publications, seminars, research, and consultations, the Council works to ensure that private investments in education produce results.

BEST COPY AVAILABLE

Testing and Evaluation

Learning from the Projects We Fund

George F. Madaus
Walt Haney
Amelia Kreitzer

Boston College

Policy Issues in the Conduct of Corporate Support of Education is a series of brief reports on topics related to the management of corporate education-support programs. They are designed to give contributions officers a more comprehensive understanding of the structural, legal, and social dimensions of corporate involvement in education support.

Copies of this report are sent free of charge to supporters of the Council. Others may purchase copies for $15 each, prepaid, from the Council for Aid to Education, Publications Department, 51 Madison Avenue, Suite 2200, New York, NY 10010. Phone 212/689-2400.

Special thanks are due to Aetna Foundation, which underwrote publication of this Policy Issues report. The opinions expressed are those of the authors and the Council, and do not necessarily reflect the opinions of Aetna Foundation or its staff.

Copyright © 1992 by the Council for Aid to Education

Table of Contents

Foreword v

Introduction 1

Chapter 1: **Metaphors and Models of Evaluation**

Metaphors Influencing Educational Evaluation 3
Three Major Approaches to Educational Evaluation 6
An Argument for Eclecticism 16

Chapter 2: **Taking an Active and Informed Role in Evaluation**

Budgeting an Evaluation 18
Evaluation Tasks 19
Tools to Help You 20
Questions You Need to Ask 27

Chapter 3: **The Role of Testing in Evaluations**

What is a Test? 29
Using Multiple Choice Tests in Evaluation 34
Using Alternative Assessments in Evaluation 37
Questions to Ask When Using Tests in Program Evaluation 40

Conclusion: Learning from Funded Projects 42

References 44

Glossary 47

Authors' Acknowledgments

The authors would like to thank Arnold Shore, president of the Council for Aid to Education, for first conceiving of this project and Gail Promboin, senior program officer of Aetna Foundation, for funding it. In addition, we gratefully acknowledge the assistance of Chris Lyons, Claudette Mungal, and Jennifer Greben at the Center for the Study of Testing, Evaluation, and Educational Policy at Boston College.

Foreword

The Council's hope in commissioning this report was to provide corporate contributions administrators with an evaluation *primer* — not on how to conduct evaluations yourself, but rather on how to be actively, purposefully, and usefully involved in initiating, monitoring, and applying the results of independent evaluations of your educational projects and programs.

George Madaus, Walt Haney, and Amelia Kreitzer have, we think, more than lived up to our expectations. Their little book can be read in an hour and used throughout a career. It will enable you to understand evaluators' perspectives on the one hand, and on the other, to ask systematic and probing questions about the choices that must be made in designing any program assessment. The chapter on educational testing sets forth the uses and limitations of both standardized and alternative forms of testing, with a degree of clarity that will surely be valuable as you make decisions about this hotly-contested aspect of educational program evaluation.

We know we must learn from the projects we fund, though it's not always obvious how to go about it. *Testing and Evaluation* should make that learning process smoother and less mysterious for us all. We extend our thanks to the authors, whose concern for improving project and program performance is genuine, heartfelt, and professional.

Arnold R. Shore
President

Introduction

Corporate support of education by any index — dollars contributed, school/business partnerships, volunteer work, equipment donated, scholarships — has grown enormously over the past two decades. By all indicators, corporate involvement in our nation's schools will continue to grow through the end of the century. Given their fiscal and human investment, and the nation's economic and civic stake in improving our schools, corporations need answers to critical questions about the projects they fund. Such questions include:

Are we making a difference?

What's working? Under what conditions? With what populations?

Are our efforts contributing to educational reform?

Are we advancing our corporate philanthropic goals?

Answers to these questions hinge on systematic evaluations of corporate-supported educational reform efforts in general and of individual projects in particular. Now more than ever, in order to maximize the impact of their limited contributions budgets, corporate funders need the skills to plan and commission appropriate project evaluations, oversee their implementation, and correctly interpret results. Since evaluations of educational programs almost always use test information as the principle measure of what students know and can do, it is imperative that corporate funders also understand the strengths and limitations of the current testing technology.

The purpose of this report is to give contributions administrators an overview of issues in program evaluation, including the place of testing in such evaluations, so that they are in a better position to plan, commission, monitor, evaluate, and use evaluations of the projects they fund. In order to accomplish these goals, we start by exploring metaphors and models for educational evaluation in the first chapter. We consider how metaphors used to describe the educative process and schools can strongly influence approaches to program evaluation. We also describe the three main approaches to educational program evaluation. In the second chapter, we offer some practical how-to advice on commissioning, monitoring, evaluating, and using evaluations. The third chapter is devoted to educational tests and issues related to their use in program evaluation.

Chapter 1

Metaphors and Models of Evaluation

Metaphors Influencing Educational Evaluation

The way we perceive education in general and schools in particular is greatly influenced by the prevailing metaphors used when discussing education. Metaphors give shape to ideas and concepts; they are linguistic tools for seeing, understanding, and experiencing one kind of thing in terms of another (Lakoff, 1980). Our metaphors and figural language influence the way we understand and talk about education; they create mind-sets and influence behavior towards schools and teachers; they also influence the kinds of questions we ask about educational programs. Most importantly, in terms of this report, metaphors influence the way we think about and design evaluations of educational programs.

Since the middle of the 19th century a consistent and powerful class of defining metaphors for schooling has been that associated with factories, engineering, assembly lines, production, or machines. The metaphor of *schools as factories* emerged from a defining social movement of the early 19th century: industrial capitalism's developing commitment to standardization, uniformity, precision, clarity, quantification, and rational tactics. This movement began with the development of the technology of interchangeable parts by the Army Ordnance Department, and eventually spread to the manufacture of a host of mass-produced products in the 20th century (Smith, 1987; Staudenmaier, 1985, 1988, 1989). The factory model with its techniques of conformity and "the perpetual supervision of behavior and tasks" (Kritzman, 1990) eventually came to be held in high esteem by educational administrators. As one observer has commented, "Like the manager of a cotton mill, the superintendent of schools could supervise employees, keep the enterprise technically up to date, and monitor the uniformity and quality of the product" (Tyack, 1974). Within this model, standardized testing became an important technique for superintendents to monitor output and hold students and educators accountable.

The factory metaphor plays out like this:

- The curriculum is the means of production.
- The student is the raw material to be transformed into a finished and useful product.
- The teacher is a highly skilled technician.

The outcomes of production are carefully plotted in advance according to rigorous design specifications.

Certain means prove wasteful and are discarded in favor of more efficient ones.

Great care is taken to see that raw materials of a particular quality or composition are channeled into the proper production system.

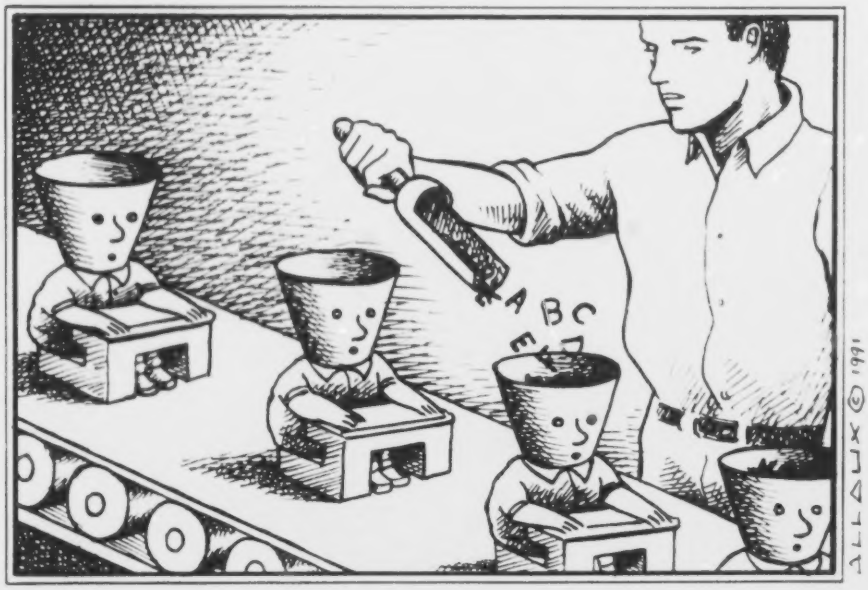
No potentially useful characteristic of the raw material is wasted.

Prospective employers are the consumers of the finished product. (Kliebard, 1972)

The factory metaphor of schooling persists to this day. For example, a recent public service advertisement in *The Atlantic Monthly* proposed replacing the "dumbed down" assembly line of the past where teachers perform routine repetitive acts, with a new kind of factory metaphor: the "high tech assembly line," where teachers trouble-shoot and problem-solve (Doyle, 1991). Not surprisingly, business and commercial terms such as outputs, bottom line, deregulate, destandardize, competition, and free market forces permeate current educational reform proposals.

Figure 1.1

The Factory Model of Education



Jean François Allaux. This illustration first appeared in *The Atlantic Monthly*, September 1991.

The factory metaphor, however, does not capture the reality of schools. Schools are not places where things are mass-produced; teachers are not assembly-line workers or robots; educated persons are not stamped out or assembled, and they do not come with warranties. Books like Tracy Kidder's *Among Schoolchildren* (1989), Philip Jackson's *Life in Classrooms* (1968), and Ted Sizer's *Horace's School* (1992) provide convincing testimony that schools are extremely complex, organic, densely-packed social systems. As one of these observers put it,

> [The teacher] is engaged in a process that is qualitatively unlike the descriptions implied in learning theories and in...the engineering view of educational progress....As typically conducted, teaching is an opportunistic process. That is to say, neither the teacher nor his students can predict with any certainty exactly what will happen next. Plans are forever going awry and unexpected opportunities for the attainment of educational goals are constantly emerging. The seasoned teacher seizes upon these opportunities and uses them to his and his student's advantage. (Jackson, 1968)

Nonetheless, the factory metaphor continues to shape the way we think about educational programs and, as we will show, the way we evaluate those programs.

As an antidote to the factory metaphor, and to illustrate how an alternative metaphor would demand a different conceptualization of schooling and the evaluation process, we might explore the metaphor of *schooling as travel*. This metaphor suggests that:

- The curriculum is a route over which students travel.
- The teacher is an experienced guide and companion.
- Each traveller will be affected differently by the journey.
- Effects are at least as much a function of the predilections, intelligence, interests, and intent of the traveller as of the contours of the route.
- This variability is not only inevitable, but wondrous and desirable.
- No effort is made to anticipate the exact nature of the effect on the traveller.
- A great effort is made to plot the route so that the journey will be as rich, as fascinating, and as memorable as possible. (Kliebard, 1972)

Two other metaphors that underlie some views of the educational process are *gardening* and *medicine*. If schools are gardens, then teachers are gardeners, whose job it is to nurture each individual student/plant by giving it individualized care (that is, the right amount of water, appropriately timed pruning, staking if necessary) and ensuring that it enjoys ideal conditions (sunlight, proper soil acidity) so that it can bloom and thrive. In marked contrast is the medical metaphor, in which the teacher/doctor diagnoses what is wrong with the student/patient, and prescribes remediating treatment/instruction.

Our views of education, embedded in our implicit and explicit metaphors of schooling, necessarily guide us in designing education support programs and choosing projects to fund. In addition, when we plan an evaluation, our views of education and those of the evaluators we hire influence our evaluation questions and criteria, our means of gathering evidence, and the nature of that evidence.

For example, consider a new district-wide whole-language program, where language arts instruction is integrated into a theme-based curriculum. Among the explicitly stated program goals are improved reading and writing skills. An evaluator whose views of the educational process are strongly colored by the factory or production metaphor will focus the evaluation effort almost exclusively on the stated bottom lines: Did reading and writing skills improve? This evaluator will strive to reach a summary judgment by first operationalizing the program goals (that is, defining the goals in measurable terms) and then measuring their degree of attainment for all students in a uniform manner. This, as we shall see in the next section of our report, is the goals-oriented approach to evaluation.

An evaluator who likens education to travel will take a different tack with the whole-language program. The view of education as travel, as explained above,

celebrates the uniqueness of each individual's educational experience and recognizes the unpredictability of the effects of education. An evaluator working from these assumptions will not focus the data collection and analysis on reaching a summary statistic about overall accomplishment of predetermined goals. Instead, the evaluator will show a concern with program processes and with differential impacts on different individuals and groups of individuals. Data will be collected in an ongoing manner, perhaps through in-depth observation or interviews. Such an evaluation might, for example, report multiple indicators of multiple program effects at different sites and for different groups. It might include a detailed case study of how the program was implemented at several sites in the district, so that a reader could understand why the program did or did not work well, with what groups, and under what conditions. As we will see below, such procedures characterize the naturalistic or responsive approach to evaluation.

Three Major Approaches to Educational Evaluation

The whole-language scenario described above suggests how variations in our assumptions about education lead to varying ways of viewing program evaluation. There are other factors as well behind the development of different concepts of evaluation. These include the rapid growth of the field since 1960; the wide range of evaluative situations and contexts; the politics and funding of evaluations; and numerous practical constraints.

In this section of the report, we explain the three most widely used approaches to program evaluation: the goals-oriented approach, the decision-oriented approach, and the responsive or naturalistic approach. (For a more detailed survey of these and other models of program evaluation, see Madaus, Stufflebeam, & Scriven, 1983.)

These three approaches can be considered as differing chiefly in the scope of evaluation inquiry, or how wide we cast the net of our evaluation questions. Goals-oriented approaches, as the name implies, consider only predetermined program goals as objects of evaluation. Decision-oriented approaches, which were developed in part to overcome the shortcomings of the goals-oriented model, expanded the objects of evaluation to include program inputs, processes, goals, and, in one specific model, context. The third approach to evaluation, variously referred to as the naturalistic or responsive approach, expanded the scope of evaluation inquiry even further. Naturalistic evaluations can explore, in fact, everything related to a program, most notably the perspectives of all the groups affected by or interested in the program. In addition, this last approach distinguishes itself from goals- and decision-oriented models in epistemology and, typically, methodology.

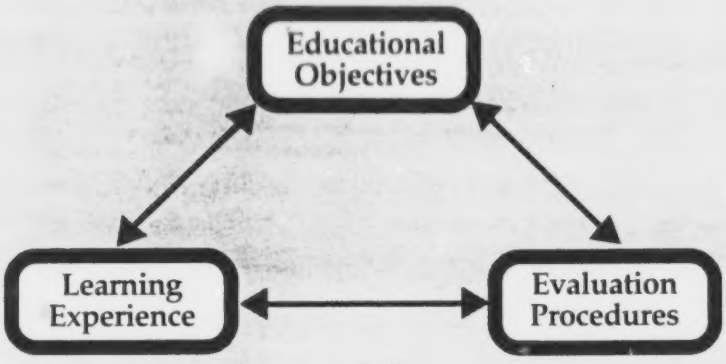
Goals-Oriented Approaches

The goals-oriented approach to educational evaluation is the oldest and perhaps most widely used model of program evaluation. It focuses on the intuitively appealing concept that the function of evaluation is to determine the extent to which an educational program has achieved its stated goals or objectives. With the focus

on outcomes — the achievement of predetermined goals or objectives — the goals-oriented approach fits comfortably within the production metaphor of schooling.

The model traces its roots to the 1930s, when Ralph Tyler conducted the famous Eight-Year Study, a longitudinal, comparative evaluation of students in thirty progressive and traditional high schools. The viability of progressive schools had become problematic when many prestigious colleges changed their entrance requirements to include courses in specific subjects like American history and physics. Progressive schools, however, were not organized around specific courses; their focus was on learner outcomes. As a result, the new college entrance criteria posed real difficulties for graduates of progressive schools. Beginning in 1934, Tyler and his associates worked with the staff and clientele of the thirty schools to help them formulate educational objectives and then assisted them in developing techniques to assess the extent to which they were successful in reaching those objectives.[1]

Figure 1.2
The Tyler Model



In a 1950 monograph, *Basic Principles of Curriculum and Instruction*, Tyler drew on the Eight-Year Study and articulated his rationale for evaluation. He saw evaluation as an integral part of the curriculum development process. He graphically portrayed the process as a triangle.

At the apex of the Tyler triangle — the beginning point of curriculum development — are educational objectives. Educational objectives are clear statements of what it is we want students to know or be able to do as a result of a program. Once the objectives are clearly delineated, the next angle of the triangle is concerned with designing and organizing the educational experiences that are likely to help students master those objectives. The final leg of the triangle is concerned with determining whether the objectives are being attained — that is, evaluating the program in terms of the objectives.[2]

In response to the Russian launch of Sputnik in 1957, the federal government began to fund large-scale curriculum development projects in math, science, and foreign languages. People naturally turned to Tyler's curriculum

[1] For a detailed description of the Eight-Year Study see Madaus & Stufflebeam, 1989; Smith & Tyler, 1942.

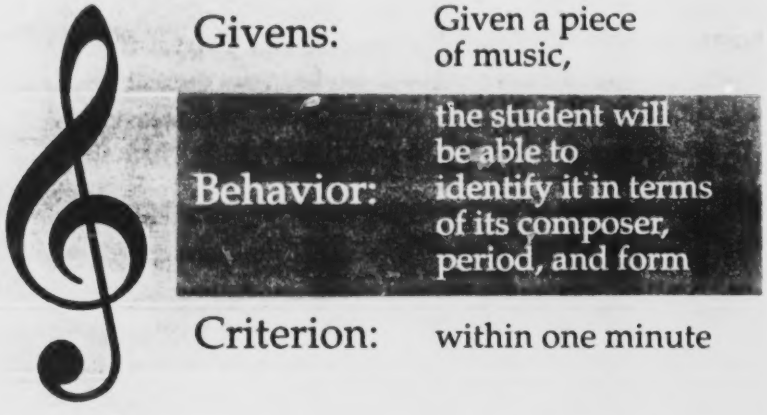
[2] Tyler's model is, of course, much more complex than this simple description of the legs of the triangle. For example, he spends considerable time discussing the sources and evaluation of the educational objectives, and the development and organization of instruction.

development approach to help them shape these new curricula. Later, when Congress asked for evaluations of the programs, a highly simplified version of Tyler's goals-oriented evaluation was frequently adopted. These goals-oriented evaluators focused exclusively on the degree of attainment of the pre-specified objectives of the program.

Figure 1.3

Delineating Objectives

Goals-oriented evaluators begin by asking for very specific statements of educational objectives in terms of student behavior and specific content. Figure 1.3 shows an example of such a statement.


Givens:
Given a piece of music,
Behavior:
the student will be able to identify it in terms of its composer, period, and form
Criterion:
within one minute

Once the objectives are explicitly delineated, the next step is to develop assessment techniques that permit students to demonstrate the behavior in question. If the objective is clearly stated, the form the assessment can take is also clear. Once measures of the objectives are developed, they are administered as a pre-test to students before the program begins. The pre-test provides a baseline against which to compare performance at the end of the program, when the students take the post-test (see figure 1.4). Changes from pre-test to post-test in the percentages of students mastering each objective become the key criteria of the program's success.

The goals-oriented approach to evaluation is appealing for many reasons:

It is good common sense to ask whether a program has met its goals. Consequently, the model is widely used and credible.

It forces program personnel to be clear about their intended outcomes and can be used to hold them accountable for the attainment of those outcomes.

It minimizes disruption and intrusion on the part of the evaluator, who only appears briefly to administer tests.

The approach is relatively inexpensive, particularly when standardized machine-scoreable tests are used.

It provides easily quantifiable, "objective" information about student performance.

The goals-oriented approach also has a number of disadvantages. The first problem with the approach became clear after the initial wave of goals-oriented evaluations. When those evaluations reported that many new programs failed to achieve their objectives, there was no way of knowing what went wrong. Similarly,

BEST COPY AVAILABLE

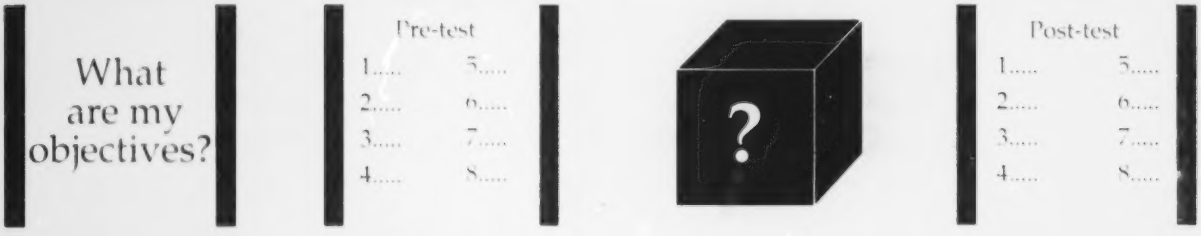
it the evaluation found a program to be successful, it could not tell educators what aspects of the program were responsible for the successes. In other words, the goals-oriented approach failed to look at resources associated with the program or at educational processes themselves. This shortcoming became known as the "black box" problem (see figure 1.4).

The second problem with the goals-oriented approach was that program goals tended to be accepted at face value and hence not subjected to critical evaluation. A program might succeed in meeting its goals, but the goals themselves might be ill-conceived or trivial.

A third problem was that by concentrating on measuring the goals of the program, unintended positive and negative outcomes went unrecognized. For example, while students may have mastered new content in a subject like mathematics, they may also have come to hate or fear the subject. The music appreciation program may have met its goals of increased student knowledge about and appreciation for music, but because the program pre-empted geography, students may have missed important content there. This lack of information about side effects made it impossible to assess opportunity costs associated with the program and to form valid judgments about the value of the program.

Figure 1.4

The Goals Oriented Approach



A final criticism of the goals-oriented approach is that it is summative in nature, not formative. A formative evaluation collects data while the program is in progress so that the information gathered can be used for program improvement. The primary audience for a formative evaluation is the program staff. A summative evaluation, on the other hand, is meant to produce data to assist funders or decision-makers in making continuation or termination judgments about a developed, fully operating program (Scriven, 1981).

Decision-Oriented Approaches

The decision-oriented approach to educational evaluation emerged in the late 1960s in response to shortcomings associated with the goals-oriented approach. One of the leading proponents of this approach defined decision-oriented evaluation as the process of delineating, obtaining, and applying descriptive and judgmental information concerning a program's merit, as revealed by its goals, design,

implementation, and results, for purposes of decision-making and accountability (Stufflebeam & Shinkfield, 1985).[3]

Many specific evaluation models fall under the decision-oriented heading; what they share are the following characteristics:

- They include as possible objects of evaluation not only program goals, but also inputs, processes, and, sometimes, context. One, some, or all of these components may be investigated in a decision-oriented evaluation.
- Theoretically, decision-oriented models are integrated into the management system of an institution (a school or school system, for example).
- Judgments about the ultimate value of the program are made not by the evaluator but by the program administration, that is, the decision-makers.
- Decision-oriented models emphasize documentation, that is, leaving a "paper trail" for accountability.

Decision-oriented evaluation models generally are concerned with three components of an educational program: inputs, processes, and outcomes or product. Hence, this class of evaluation models is sometimes referred to as the IPO approach. As with goals-oriented models, and as is evident from the terminology, an implicit production metaphor is inherent in the IPO approaches.

- **Inputs** are resources, such as money, staff, facilities, books, or computers, allocated to an educational program. Evaluations of inputs most often involve compiling an inventory of resources and checking whether resources available comply with those specified in program plans.
- **Processes** refer to the way resources are actually utilized in the program itself. Examples of process questions are: "What's happening in the program?" "How are inputs experienced by program participants?" and "What is the nature of the educational experience provided?" Process evaluation attempts to look inside the black box in order to better understand the mechanisms that produce program outcomes.
- **Outcomes** or **products** are what the program produces. Outcome evaluations look at the extent to which a program has attained its stated goals. It has all the characteristics of the goals-oriented approach discussed above. However, most decision-oriented outcome evaluations are also concerned with assessing side-effects or unanticipated outcomes.

Some decision-oriented evaluation models also consider the context of a program (Madaus et al., 1983).

- **Context** refers to the larger political and situational context in which the program is embedded. Context evaluation was added so that the choice of program

[3] Other important theorists of decision-oriented models include Provus, 1971; Cooley & Bickell, 1986; and Brinkerhoff, 1983.

goals could be properly evaluated. Context evaluations can involve conducting a needs assessment to help in the formulation or evaluation of program goals (see boxed insert below).

Context, inputs, process, and products can singularly or together (as in the CIPP approach) be the object(s) of an educational evaluation; the information needs of the decision-maker(s) determine which of the objects becomes the focus of the evaluation. As table 1.1 makes clear (see page 12), different objects lead to different evaluation questions, procedures, and outcomes.

A Closer Look

Needs Assessment

A needs assessment is an evaluation that compares the "is" of a system with the "should be." Typically, evaluators collect and interpret data about the present status of a system and identify discrepancies between what is and what is desired. These data are then used to evaluate, reformulate, and set program goals, and to assign priorities among those goals. Some of the contextual questions that might be considered in a needs assessment are:

- What are the predictable problems?
- How embedded are they in the history, tradition, and organization of school systems?
- What courses of preventive action might be employed, at least to minimize these problems?
- When and why should we expect which outcomes? (Sarason, 1990)

The advantages of the decision-oriented approach are:

- It can help evaluators and managers generate questions about a program.
- It is adaptable to a wide range of evaluative purposes, but also puts any particular evaluation into sharp focus by defining the essential information needed for a decision.
- It stresses the utility and timeliness of the information collected.
- It can serve program improvement.
- It can look for unintended outcomes.
- It is integrated into the decision-making apparatus of an organization.
- The goals themselves can be subject to evaluation.
- Like goals-oriented models, it typically provides easily quantifiable, "objective" information about student performance. (Worthen & Sanders, 1987)

The disadvantages of decision-oriented models are:

- The information needs of the decision-makers, such as program managers, drive the evaluation. Thus, the managers' concerns and focus can limit the scope of the evaluation so that the information needs of other audiences (teachers, parents, the public) may not be accommodated.

Table 1.1

Four Objects of Evaluation

	Context Evaluation	Input Evaluation	Process Evaluation	Product Evaluation
Information Needs of Decision-Maker	Guidance for choice of objectives & assignment of priorities	Guidance for implementing and refining program	Guidance for implementing and refining program	Guidance for termination, continuation, or modification of program
Questions Informed by the Evaluations	• What is the institutional context? • What are the target populations & their needs? • How can needs be addressed? • What are the problems behind those needs? • Can proposed objectives respond sufficiently to those needs?	• What are system capabilities? • What are possible program strategies & procedural designs for implementing those strategies? • What's the budget? • What are time constraints?	• How is the program working? • What defects are turning up or are likely to turn up in procedural design or its implementation? • What specifically is working well & for whom? • Is program proceding according to plan?	• What were outcomes? • What were stakeholders' judgments of those outcomes? • How do outcomes relate to objectives? • How do outcomes relate to context, input, & process information? • What are the worth & merit?
Methods/Activities	• system analysis • surveys • document review • hearings • interviews • diagnostic tests • consensus-building techniques	• analysis of available human & material resources • inventory of possible strategies • inventory of possible procedural designs • cost & feasibility analyses • literature search • visits to exemplary programs • pilot trials	• monitoring problems • remaining alert to potential & unanticipated problems • obtaining specified information for program decisions • describing the actual process • ongoing staff interviews • ongoing program observation	• defining goals operationally • measuring outcome criteria • collecting judgements of outcomes from stakeholders • quantitative analyses • qualitative analyses
Documentation Collected or Produced for Accountability	• record of objectives and bases for their choice • record of needs, opportunities & problems	• record of chosen strategy & design • rationale for choice	• record / log of actual program activities and process	• documentation of actual outcomes (intended & unintended, positive & negative)

Madaus, G., Scriven, M., Stufflebeam, D. (1983). *Evaluation models: Viewpoints on educational and human services evaluation.* Boston: Kluwer-Nijhoff Publishing. Used with permission of the publisher.

1. It can be costly and complex if priorities are not clearly set.
2. It may not be practical or affordable for small or poorer school systems.
3. It assumes that important decisions can be clearly identified in advance, that the decision-making process is orderly, predictable, and rational, and that data will actually be used to inform decision-making. (Worthen & Sanders, 1987)

Responsive or Naturalistic Approaches

Around 1970 a backlash emerged against the dominance of goals-oriented and decision-oriented approaches to educational evaluation. Critics pointed out that these quantitative approaches failed to capture what was really going on in many innovative programs, and the reports generated failed to give people working in the programs the kind of information they needed to better understand and reflect on what they were doing (Worthen & Sanders, 1987). This criticism of the two traditional evaluation approaches went far beyond methodological considerations of paper-and-pencil testing versus direct observation in classrooms; it reflected instead basic epistemological differences on how best to seek "truth" in the human experience; it involved questions of what kind of data should be gathered and for what reasons; it involved considerations of whether to pursue a holistic or an atomistic approach to trying to understand the nature of a program. The disquiet felt with the traditional approaches to evaluation is captured in the following observation by Lewis Thomas about the physical sciences (which stongly influenced the quantitative mind-set of educational evaluation).

> Science gets most of its information by the process of reductionism, exploring the details, then the details of the details, until the smallest bits of the structure, or the smallest parts of the mechanism, are laid out for counting and scrutiny. Only then when this is done can the investigation be extended to encompass the whole organism or the entire system....Sometimes it seems that we take a loss, working this way. Much of today's public anxiety about science is the apprehension that we may forever be overlooking the whole by an endless, obsessive preoccupation with the parts. (Thomas, 1980)

Dissatisfaction with the quantitative and mechanistic approaches to evaluation — with their emphasis on using paper-and-pencil techniques to measure operationally-defined objects and the decomposition of programs into inputs, processes, and outcomes — eventually spawned a number of new approaches to evaluation which are generally referred to as naturalistic or responsive approaches. In effect, proponents of the naturalistic approach to evaluation start out from a different set of metaphors for schooling than do traditionalists. Rather than seeing schools as production lines, they would be more comfortable with the travel metaphor described above.

The naturalistic approach to evaluation includes many different models, such as the case study method and the illuminative model, but all models share some features.[4]

Naturalistic evaluations center on activities, transactions, and effects occurring within the program rather than on program goals. The focus of such evaluations is not determined in advance, as it is in goals- and decision-oriented models (which have, for that reason, been termed "preordinate" models). Instead, the evaluation design and focus emerge through observation of program transactions; eventually, through an iterative process, themes or issues surface that the evaluator and other interested parties agree should be zeroed in on. Naturalistic models of evaluation develop evidence on both sides of an issue and are sensitive and responsive to different information needs and to the value positions of different audiences. Evaluations in the naturalistic tradition attempt to describe the characteristics and nature of engagements between program participants — students and teachers, for example. They strive to provide a rich, thick, illuminating, comprehensive, qualitative description of the program — a description aimed at awakening audiences to unseen dimensions of the program. Their intent is to illuminate and clarify what is actually going on in the program so that program participants can discern and reflect on what they are doing (see box opposite).

Methods employed by naturalistic evaluators differ from the quantitative, statistical methods typically used in goals- and decision-oriented evaluations. Naturalistic evaluators use techniques that have long been associated with disciplines such as anthropology: long-term, direct observation; open-ended interviews; document or artifact analysis; and in-depth case studies. Indirect paper-and-pencil data-collection is not taboo, but might be employed to confirm conclusions drawn from qualitative data.

Naturalistic evaluation and preordinate models of evaluation are contrasted in table 1.2 (see page 16) across eleven dimensions.

The advantages of responsive/naturalistic approaches to evaluation are:

- They are problem- and issue-oriented.
- They provide thick, rich descriptions of what goes on in a program.
- They accommodate the information needs of multiple stakeholders.
- They illuminate and help people reflect on what is happening in a program.
- They are excellent vehicles for formative evaluation directed at program improvement.

The disadvantages of responsive/naturalistic approaches to evaluation are:

- They are time-consuming and labor-intensive and therefore more costly than other approaches.

[4]Prominent models and their proponents include: Stake, 1983a; Stake, 1983b; Guba & Lincoln, 1981, 1983; Eisner, 1983, 1991; Denny, 1978.

A Closer Look

Quantitative and Qualitative Data

Below are two examples of data reporting. The first is an example of how teacher behavior might be described in a report that collected quantitative data. The second typifies qualitative data reporting in its use of "thick, rich description" to convey a sense of teacher behavior. Of course, both are excerpted and out of context here.

Example 1: Quantitative

Launching is primarily a function of the teacher rather than the pupil. The teacher speaks 85.2 percent of all structuring components; in contrast, the pupil speaks only 10.2 percent. The range for the teacher is from a high of 96.8 percent in Class 6 to a low of 73.5 percent in Class 13, with a median of 86.2 percent. The range for the pupil, on the other hand, is from 23.6 percent in Class 13 to zero in Classes 9 and 10, with a median percentage of 11.8 percent. (Eisner, 1985)

Example 2: Qualitative

The first few bars from George Winston's *Autumn* gently remind Ms. Morrow's third-graders to retrieve their book tubs and find a niche for private reading time. Ready for their weekly conferences, Rachel, Towanda, Juan, Marcus, and Chi Li claim the spot under the table today, leaving the others room to spread out along the table. While the children shuffle through their books and reading journals, Ms. Morrow settles into the empty seat next to Marcus.

"Me first today?" his plaintive tone a thin mask for his pride and eagerness, "I'm almost done — chapter nine!"....

"Good for you! Just let me remind the others of what they need to do. Towanda and Chi Li, Juan and Rachel, remember to write your weekly letter sometime today — either now before our conference or during buddy reading time. I like how you have been describing your books and what happens — lots of colorful words — so keep up the good work! Remember to think and write about your reactions to the book, too." After making their decisions, the four children quickly become engrossed in their work.

Perched on the edge of his seat, Marcus squirms and sighs expectantly, barely containing his impatience to start his conference. Ms. Morrow invites him to bring her up to date on his current book while she retrieves a conference form out of his reading folder. She listens intently to his description, noting on the conference form "Emphasizes characters, plot, & his own reaction to book."

As Marcus finishes, Ms. Morrow follows up with "Did the main character change during the story?" A ponderous look appears on his face before he nods, "How?" And for the next four or five minutes their conference continues until Marcus beseeches, "Let me read you this one part — it's my favorite so far." While he searches for page and paragraph, Ms. Morrow jots down snippets from their conference on the form:

Date	Book	Comprehension & Strategies
11/30	Top Secret	• he asked for Chris' recommendation • said I couldn't remember much about bk, he said "Read the back cover!" • previewed book together • ENTHUSIASTIC!!
12/6	Top Secret	• excited to tell me about bk • emphasizes charac/plot/own reaction • knew how character had changed • couldn't tell how change affected plot • wanted to read aloud fav part

Ellwein & Graue, reprinted with permission from Macmillan Publishing Company from their upcoming book *Camps and Classroom: Making Education Multicultural*, edited by C. Grant and M.L. Gomez. All rights reserved.

- The results are not generalizable in the traditional, "scientific" way.
- They are not a credible model for many people who fund evaluations and who have a different epistemological view of "truth." The results are viewed by some as "subjective" or "soft" rather than as "objective" and "hard."

Table 1.2
Preordinate vs. Responsive Approaches to Evaluation

	Preordinate Approaches Goals-Oriented & Decision-Oriented Models	Responsive / Naturalistic Approaches
Purpose	to determine the degree to which goals are achieved	to help audiences discern program strengths & weaknesses
Audience	client	stakeholders
Scope of Services	information requirements agreed upon at outset	responds to information requirements of audiences as study unfolds
Contracts	obligations defined & negotiated at beginning	general purpose outlined initially; plan becomes more specified over time
Main Orientation	intents & indicators of those events	issues/events: what do certain audiences want addressed?
Evaluation Design	pre-specified	emergent
Methodology	scientific research model	to reflect to people what they're doing
Techniques	experimental design, hypothesis testing, statistics, objective tests	case studies, expressive reports
Evaluation / Client Communication	formal & infrequent	informal & continuous
Basis for Valuation Interpretations	pre-stated objectives, norm groups, competing programs	differs depending on value perspectives of stakeholders
Provisions for Reducing Bias	experimental controls	replication, seeking confirming evidence

Stufflebeam, Daniel L. & Shinkfield, Anthony J. (1985). *Systematic evaluation: A self-instructional guide to theory and practice.* Boston: Kluwer-Nijhoff. Used with permission of the publisher.

An Argument for Eclecticism

Which model should we choose for the evaluation of our education projects? The emphatic answer is, "None of the above; that's the wrong question to ask." Instead of starting from a pet approach, we should begin with a consideration of the evaluation questions that could be addressed, the issues that must be addressed, and the resources available for the evaluation. Each evaluation approach discussed above has its particular strengths that can help illuminate different aspects of a program.

Within the limitations of your budget, pick and choose features from various models that can provide the best evidence to answer your questions about the project. For example, you might want to consider combining testing data from a goals-oriented approach, resource allocation data from the decision-oriented approach, and observational and interview data from the naturalistic approach when evaluating a project.

If an evaluator you are considering hiring is known to espouse a particular model, explore the possibility with him or her of differentiated staffing to embrace people with other evaluative viewpoints, skills, and orientations. Assembling an evaluation team with different skills and approaches is one way of ensuring that the evidence collected to describe or judge your program is broad-based and relevant, and addresses the different information needs and value perspectives of various concerned audiences.

Chapter 2

Taking an Active and Informed Role in Evaluation

Becoming an expert in formal evaluation is not something that can be done overnight, nor can the full repertoire of evaluation skills be covered in a few pages. This chapter, therefore, is written with the assumption that you, as an education funder, are committed to formal, independent evaluations of your projects. Even hiring an outside evaluator, however, entails major responsibilities for the corporate contributions staff. You need to become familiar with key evaluation concepts and tools so you can take an active and informed role in the evaluation of projects. Becoming handy with the tools and concepts will enable you to:

- ask appropriate questions of the people who evaluate your programs,
- evaluate the evaluations you receive, and
- help ensure that the evaluation will be used.

This chapter opens with a discussion of evaluation budgets. Next, we outline the tasks that are common to any evaluation. Finally, we introduce two tools that can help you in planning, monitoring, and using evaluations.

Budgeting an Evaluation

Budgets are always a factor in deciding how to craft or combine components from different models in planning your evaluation. There is no fixed rule for determining the size of an evaluation budget; it depends on what you need to find out. A very crude rule of thumb is that an evaluation budget is between 3 and 10 percent of the project budget. A moment's reflection on this rule of thumb reveals that the size of the program budget can seriously limit your evaluation options. A project funded at the million dollar level would provide roughly $30,000 to $100,000 for evaluation, giving you the flexibility of combining features from various models. But if a project is funded at the $20,000 level and the rule of thumb applied, the $600 to $2,000 available for evaluation would not support a labor-intensive field-study component. In evaluations, as in life, "you get what you pay for."[5]

The dollar costs of evaluations can be grouped into categories as in table 2.1 (see page 19).

[5] For tips that might help minimize evaluation costs in some situations see Fink & Kosecoff, 1983.

The estimates in table 2.1 are admittedly crude. Costs can vary widely in some categories. For example, one review found professional staff costs to range from 52 percent to 90 percent of the total budget, and data processing costs from 0 to 23 percent (Alkin & Stecher, 1983). Nonetheless, these categories and percentage estimates of total costs should provide rough baselines when reviewing any budget submitted in an evaluation proposal.

Table 2.1
Categories of Evaluation Expenses[a]

Category	Share of Evaluation Budget
Professional evaluation staff salary and benefits	70%
Clerical/secretarial staff	16%
Consultants	2%
Travel and per diem for staff and consultants	3%
Communications (e.g., fax, phones, postage) Printing and duplication Printed materials (e.g., tests, instruments, books, journals) Office supplies	5%
Data processing (e.g., data coding, checking and entry, programming, computer time)[b]	3%
Subcontracts: Includes any costs for support services such as repair of equipment, accounting, legal services.	varies
Overhead (e.g., costs for facilities, equipment, utilities). This highly variable cost is usually a fixed percent of the cost of the evaluation.	varies

Some people argue that money spent on evaluation is money that could be better spent for direct program services. But this view is shortsighted. There can be serious fiscal and opportunity costs associated with a decision not to do an evaluation. An evaluation done correctly should uncover strengths, weaknesses, cost benefits, and opportunity costs; it should also contribute to program improvement. In fact, evaluations, if properly designed and used, should be "cost-free" (Scriven, 1981). That is, you should be able to show a net positive return over time for the evaluations you fund; they should on average more than pay for themselves in terms of program improvement.

Evaluation Tasks

The previous chapter sketched out the features of three different approaches to evaluation and argued that eclecticism is preferable to orthodoxy when designing an evaluation. The approach taken depends, of course, on the purposes and context of the evaluation. Regardless of the specific model or the features of specific models that you adopt, there are certain tasks that need to be undertaken in any evaluation. These tasks can be conceptualized as belonging to three major phases: planning, managing, and using the evaluation (Joint Committee on Standards for Educational Evaluation, 1981). Typically, the corporate contributions staff is integrally involved in the first and last phases of evaluation, but the staff must also be aware of decisions being made by the evaluator in the middle phase.

[a]The cost categories for this table are drawn from Sanders, 1983, and Alkin & Stecher, 1983. Alkin and Stecher, using data from an informal survey of evaluation professionals, arrived at the estimates here regarding percent of evaluation budget.

[b]These costs are more heavily associated with the goals-oriented and decision-oriented approaches. The cost of analyzing qualitative, observational, and interview data in naturalistic evaluations is generally covered in staff costs.

In the first phase of an evaluation, the planning phase, there are three tasks, all of which demand the active participation of the corporate contributions staff:

- deciding whether or not to do an evaluation,
- clarifying and assessing the purposes of the evaluation, and
- ensuring the political viability of the evaluation study.

This last task, ensuring the political viability of your study, might not be the first concern that comes to mind when planning an evaluation. However, failure to recognize the essentially political nature of evaluations often results in evaluations that are not used. The political context of the evaluation must be identified at the outset, in what is called the front-end analysis — the preliminary research (interviews, document review, etc.) that guides evaluation design. Ensuring political viability may be accomplished in part by identifying the key audiences and their information needs.

The second phase of the evaluation, the management phase, involves five tasks, only the first of which typically entails the direct participation of the corporate contributions staff:

- drawing up the contract,
- staffing the study,
- managing the study,
- collecting data, and
- analyzing data.

The final phase of evaluation concerns the use of the evaluation, and involves two tasks:

- reporting the findings, and
- applying results.

Although the evaluator writes the evaluation report, the corporate contributions staff may decide how and to whom the disseminated findings will be disseminated (via reports, meetings, press releases, etc.). The final task in evaluation, applying the results, is the most important, and should be the ultimate focus of all prior evaluation tasks. Do the findings inform education funding and program decisions? Does the evaluation contribute to broader discussions of educational priorities? Thoughtful planning and implementation of the evaluation help to ensure that the answer to both questions can be "yes."

Tools to Help You

There are many published reports and studies on evaluation and systems of evaluation. Two guides that the authors have found particularly useful will be discussed here. They are, first, the *Standards for Evaluations of Educational Programs, Projects, and Materials*, and second, the effectiveness criteria used by the Program Effective-

ness Panel, which is the Department of Education's primary institutional mechanism for validating education programs.*

The Standards

The *Standards for Evaluations of Educational Programs, Projects, and Materials* (Joint Committee, 1981), hereafter referred to as the *Standards*, are widely recognized guidelines that can be used in all phases of an evaluation. They were developed over a seven-year period with input from more than 200 people working in the field and with the support and participation of education organizations, foundations, federal agencies, and professional evaluators. The committee that drew up the *Standards* represented a wide range of interests, including technically-oriented groups that design and carry out evaluations, practitioner-oriented groups whose members are affected by evaluations, and consumer organizations that commission evaluations and use their results. The *Standards* are designed to be "model free," in that they apply to all evaluation approaches. The strengths of the standards are that they provide:

- common language that facilitates communication and collaboration in evaluation;
- a set of general principles for dealing with a variety of evaluation problems;
- a conceptual framework by which to study various evaluation models;
- criteria by which to evaluate evaluations and hold evaluators accountable;
- a set of working definitions to guide the evaluation process; and
- a non-technical, jargon-free, easy-to-read framework for consumers who must commission and use evaluations.

The thirty standards are grouped according to four attributes associated with a well-conceived evaluation: utility, feasibility, propriety, and accuracy.

- The eight *utility* standards deal with identifying the various audiences who have a stake in the evaluation, ascertaining their information needs, and reporting relevant information to them in a clear and timely manner.
- Three *feasibility* standards are directed at ensuring that evaluations are cost-effective, efficient, practicable, politically viable, and not excessively disrupting.
- Eight *propriety* standards are concerned with legal and ethical evaluation procedures. They are meant to ensure that the rights of persons affected by an evaluation are protected.
- The eleven *accuracy* standards address the technical adequacy of an evaluation.

*Another useful evaluation resource is the Program Evaluation Kit from Sage Publications (Herman, Morris, & Fitzgibbon, 1987).

A summary statement of each standard appears in table 2.2. In the *Standards* volume, every standard is described according to the following format: 1) a statement of the standard; 2) an overview of the standard, including its rationale, definition of key terms, and explanation; 3) guidelines, that is, procedural suggestions intended to help meet the requirements of the standard; 4) pitfalls, a list of difficulties associated with the standard; 5) caveats, which concern tradeoffs that may be necessary in applying the standard in different circumstances; and 6) an illustrative case that demonstrates how the standard might be applied in context.

The *Standards* are not mechanical rules to be applied automatically across all evaluations. Instead, they are principles or guidelines that should be carefully considered in relation to key evaluation tasks. Table 2.3 (see page 26) shows how each of the standards can inform the various evaluation tasks. For example, notice that standard A1, Audience Identification, is relevant not only to the first three evaluation tasks, but also when it comes to reporting findings and applying the results at the end of the evaluation.

Table 2.2

Summary of the Standards

A. UTILITY STANDARDS

These utility standards are intended to ensure that an evaluation will serve the practical information needs of given audiences.

These standards are:

A1 Audience Identification
Audiences involved in or affected by the evaluation should be identified, so that their needs can be addressed.

A2 Evaluator Credibility
The persons conducting the evaluation should be both trustworthy and competent to perform the evaluation, so that their findings achieve maximum credibility and acceptance.

A3 Information Scope and Selection
Information collected should be of such scope and selected in such ways as to address pertinent questions about the object of the evaluation and be responsive to the needs and interests of specified audiences.

A4 Valuational Interpretation
The perspectives, procedures and rationale used to interpret the findings should be carefully described, so that the bases for value judgements are clear.

A5 Report Clarity
The evaluation report should describe the object being evaluated and its context, and the purposes, procedures and findings of the evaluation, so that the audiences will readily understand what was done, why it was done, what information was obtained, what conclusions were drawn and what recommendations were made.

A6 Report Dissemination
Evaluation findings should be disseminated to clients and other right-to-know audiences, so that they can assess and use the findings.

A7 Report Timeliness
Release of the reports should be timely, so that audiences can best use the reported information.

A8 Evaluation Impact
Evaluations should be planned and conducted in ways that encourage follow-through by members and audiences.

This matrix, together with a copy of the *Standards*, should provide excellent guidance for funders considering issues related to planning and commissioning an evaluation and choosing an evaluator. Once you have awarded a contract, the *Standards* can also assist you in monitoring the evaluator's work and evaluating and using the final report.

The PEP Criteria

The second set of tools we recommend for use in evaluation are the criteria of effectiveness developed by the Department of Education's Program Effectiveness Panel, or PEP (Ralph & Dwyer, 1988). The PEP is a national panel with over sixty members whose job is to validate or certify the effectiveness of educational programs. With the PEP stamp of approval, programs can be included in the National Diffusion Network (NDN), a program funded by the federal government for over twenty years as a way to disseminate exemplary projects and products. Even if you do not target inclusion in the NDN as one of your own eventual program goals, the

Table 2.2
(continued)

B. FEASIBILITY STANDARDS

The feasibility standards are intended to ensure that an evaluation will be realistic, prudent, diplomatic, and frugal.

These standards are:

B1 Practical Procedures
The evaluation procedures should be practical, so that disruption is kept to a minimum and needed information can be obtained.

B2 Political Viability
The evaluation should be planned and conducted with anticipation of the different positions of various interest groups, so that their cooperation may be obtained, and so that possible attempts by any of these groups to curtail evaluation operations or to bias or misapply the results can be averted or counteracted.

B3 Cost Effectiveness
The evaluation should produce information of sufficient value to justify the resources expended.

C. PROPRIETY STANDARDS

The propriety standards are intended to ensure that an evaluation will be conducted legally, ethically, and with due regard for the welfare of those involved in the evaluation, as well as those affected by its results.

These standards are:

C1 Formal Obligation
Obligations of the formal parties to an evaluation (what is to be done, how, by whom, when) should be agreed to in writing, so that these parties are obligated to adhere to all conditions of the agreement or formally to renegotiate it.

C2 Conflict of Interest
Conflict of interest, frequently unavoidable, should be dealt with openly and honestly, so that it does not compromise the evaluation process and results.

C3 Full and Frank Disclosure
Oral and written evaluation reports should be open, direct, and honest in their disclosure of pertinent findings, including the limitations of the evaluation.

(continued)

PEP guidelines for credible, persuasive evaluation evidence can be useful in designing an evaluation.

The PEP effectiveness criteria address three broad questions:

- Is the evaluation design credible?
- Are the results of the program meaningful?
- Does the object of evaluation (the program or product) have the potential to be replicated?

To answer the question concerning evaluation design, the PEP requires that an evaluation employ appropriate measurement (that is, instruments that are in line with the program's goals), technically strong measurement techniques, and careful, well-documented data collection. In addition, the evaluation must be able to link obtained outcomes with the program itself. In other words, alternate explanations

Table 2.2
(continued)

C4 Public's Right to Know
The formal parties to an evaluation should respect and assure the public's right to know, within the limits of other related principles and statutes, such as those dealing with public safety and the right to privacy.

C5 Rights of Human Subjects
Evaluations should be designed and conducted so that the rights and welfare of the human subjects are respected and protected.

C6 Human Interactions
Evaluators should respect human dignity and worth in their interactions with other persons associated with an evaluation.

C7 Balanced Reporting
The evaluation should be complete and fair in its presentation of strengths and weakness of the object under investigation, so that strengths can be built upon and problem areas addressed.

C8 Fiscal Responsibility
The evaluator's allocation and expenditure of resources should reflect sound accountability procedures and otherwise be prudent and ethically responsible.

D. ACCURACY STANDARDS

The accuracy standards are intended to ensure that an evaluation will reveal and convey technically adequate information about the features of the object being studied that determine its worth or merit.

The standards are:

D1 Object Identification
The object of the evaluation (program, projects, material) should be sufficiently examined, so that the form(s) of the object being considered in the evaluation can be clearly identified.

D2 Context Analysis
The context in which the program, project, or material exists should be examined in enough detail, so that its likely influences on the object can be identified.

D3 Described Purposes and Procedures
The purposes and procedures of the evaluation should be monitored and described in enough detail, so that they can be identified and assessed.

of results must be addressed and ruled out. Further evidence of a credible evaluation design is provided by a comparison standard, usually a carefully composed control group or appropriate norm groups.*

The PEP considers program results to be meaningful when educationally important goals are targeted and progress toward those goals is strong. Meaningful results can be demonstrated in an evaluation by establishing the need for the program and clearly stating program purposes. The PEP also recommends that evaluations specifically address the distinctiveness and superiority of the program, as compared to similar projects.

The potential for replication of a project can be established by documenting the program's generalizability and efficiency. Generalizability can be demonstrated by collecting project data from several sites or at least from several different groups over several years. If the program proves successful with different students or at different schools or districts, the transportability of the program to other locales is

*See chapter 3 for a discussion of norms and norming groups.

Table 2.2
(continued)

D4 Defensible Information Sources
The sources of information should be described in enough detail, so that the adequacy of the information can be assessed.

D5 Valid Measurement
The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the interpretation arrived at is valid for the given use.

D6 Reliable Measurement
The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the information obtained is sufficiently reliable for the intended use.

D7 Systematic Data Control
The data collected, processed, and reported in an evaluation should be reviewed and corrected, so that the results of the evaluation will not be flawed.

D8 Analysis of Quantitative Information
Quantitative information in an evaluation should be reviewed and corrected, so that the results of the evaluation will not be flawed.

D9 Analysis of Qualitative Information
Qualitative information in an evaluation should be appropriately and systematically analyzed to ensure supportable interpretations.

D10 Justified Conclusions
The conclusions reached in an evaluation should be explicitly justified, so that the audiences can assess them.

D11 Objective Reporting
The evaluation procedures should provide safeguards to protect the evaluation findings and reports against distortion by the personal feelings and biases of any party to the evaluation.

Reprinted from *Standards for Evaluations of Educational Programs, Projects, and Materials* with the permission of The Joint Committee on Standards for Educational Evaluation.

Table 2.3

Relative Importance of Thirty Standards in Performing Ten Tasks in an Evaluation

	Decide Whether to do Study	Clarify & Assess Purpose	Ensure Political Viability	Contract	Staff the Study	Manage the Study	Collect Data	Analyze Data	Report Findings	Apply Results
A1 Audience Identification	●	●	●	●		●			●	●
A2 Evaluator Credibility	●		●	●	●	●	●			
A3 Information Scope and Selection				●			●		●	
A4 Valuational Interpretation		●	●				●	●	●	●
A5 Report Clarity						●			●	
A6 Report Dissemination			●	●					●	●
A7 Report Timeliness				●					●	
A8 Evaluation Impact	●	●	●						●	●
B1 Practical Procedures			●				●	●		
B2 Political Viability	●		●	●	●	●	●			●
B3 Cost Effectiveness	●	●				●				
C1 Formal Obligation	●		●	●		●	●			●
C2 Conflict of Interest	●	●	●	●	●	●				●
C3 Full and Frank Disclosure				●					●	
C4 Public's Right to Know			●	●					●	●
C5 Rights of Human Subjects			●	●		●	●			●
C6 Human Interactions			●			●	●			
C7 Balanced Reporting							●		●	●
C8 Fiscal Responsibility			●	●		●				
D1 Object Identification	●	●		●			●	●	●	
D2 Context Analysis	●	●					●	●	●	●
D3 Described Purposes and Procedures	●	●	●	●		●	●		●	●
D4 Defensible Information Sources			●				●		●	
D5 Valid Measurement							●			
D6 Reliable Measurement							●			
D7 Systematic Data Control						●	●			
D8 Quantitative Analysis								●		
D9 Qualitative Analysis								●		
D10 Justified Conclusions		●						●	●	●
D11 Objective Reporting			●		●				●	

Madaus, G., Scriven, M., Stufflebeam, D. (1983). *Evaluation models: Viewpoints on educational and human services evaluation.* Boston: Kluwer-Nijhoff Publishing. Used with permission of the publisher.

supported. Evidence of efficiency includes data about the time, money, and effort necessary for the program, as well as comparative information about the costs of similar programs.

In addition to the general criteria summarized above, the PEP guidelines include a more detailed look at four specific types of evaluation claims.

1. Claim Type 1: Academic achievement — changes in knowledge and skills.
2. Claim Type 2: Improvements in teachers' attitudes and behaviors.
3. Claim Type 3: Improvements in students' attitudes and behaviors.
4. Claim Type 4: Improvements in instructional practices and procedures. (Ralph & Dwyer, 1988)

For each of these claim types, the PEP guidelines provide 1) examples of the claim, 2) discussion of evaluation design, 3) questions to ask about design, 4) discussion of instruments, procedures, and data collection, 5) questions to ask about instruments, procedures, and data collection, 6) discussion of results, and 7) questions to ask about results. Excerpted in the box on page 28 are some of the specifics about Claim Types 1 and 4.

Questions You Need to Ask

The expertise for which we pay independent evaluators consists in part of the technical knowledge that comes into play in the management and implementation of data collection and data analysis. The involvement of corporate contributions officers, as noted in the section above, is most extensive in the planning phase and in using the evaluation. What follows are some of the most important questions you need to ask about those two phases.

- Why evaluate? What can we expect to learn? Does the evaluation design maximize our chances of finding out what we need to know?
- Who are the relevant audiences? Does the evaluation design speak to the information needs of multiple audiences?
- Will the evaluator be credible to relevant audiences?
- Does the design allow for collection of data from different sources in different ways?
- Do the evaluation plan and the budget both include provisions for reporting to different audiences, in ways appropriate to each?.
- How can we include the results of the evaluation in relevant bodies of knowledge and collections of research?

A Closer Look

Excerpted PEP Guidelines Concerning Claim Types 1 and 4

Claim Type 1: Academic Achievement

Examples of claims:

Acquisition of factual knowledge: Students in the physics project at three typical high schools made greater gains than the national norm group on a standardized test of physics knowledge.

Application of knowledge: In addition to making greater-than-expected gains in library reference skills on a standardized test, program students required less assistance with research activities than comparison students, as measured by structured observations in school libraries.

Questions about evaluation design:

- Is this the strongest and most appropriate research design that could be undertaken? If not, what are the reasons for not choosing another design?
- How was the comparison group chosen? Is there evidence that it is similar to the project group in educationally relevant ways?

Questions about instruments, procedures, and data collection:

- Is there evidence that the instruments or procedures are reliable?
- Are the measures accepted in the field? If they are not well known, why were they selected or how were they developed?

Questions about analysis and discussion of results:

- Do the results show statistical significance?
- Are the effect sizes large enough to have practical significance? Are they plausible?
- Are there negative or positive effects on learning in subject areas unrelated to the submission claims? For example, was less time spent on other subjects to allow intensive treatment in the project area?

Claim Type 4: Improvements in Instructional Practices and Procedures

Examples of claims:

Reductions in costs and improvements in efficiency of service delivery: By instituting a cooperative program among three school districts, duplicate arts education efforts were eliminated and expanded opportunities were provided at significantly reduced costs.

Improvements in service to particular client groups: Through the satellite videoscience service, children at 600 sites participated in the master teacher science lesson series for a cost of approximately $1.00 per student.

Questions about evaluation design:

- Does the design allow for pre/post assessment or time-series assessment over a time period sufficiently long to assure stability?
- Is a variety of measures and comparisons used?

Questions about instruments, procedures, and data collection:

- If existing records are used, what safeguards exist to ensure accurate completion and maintenance of recorded information?
- If self-reporting measures were used, what was done to promote objectivity? What was done to deal with nonresponse?

Questions about analysis and discussion of results:

- Is the demonstrated change worthwhile in terms of cost? Are the time savings worthwhile?
- Do the changes or improvements compare favorably to standard practices in similar institutions?

Excerpted from Ralph, L. & Dwyer, C. M. (1988). *Making the case: Evidence of program effectiveness in schools and classrooms*. Washington, DC: US Government Printing Office. Used by permission of the authors.

Chapter 3

The Role of Testing in Evaluations

Testing is closely tied to evaluation. Tests of some sort play a role in virtually all educational program evaluations; indeed, too often an "evaluation" is no more than a hasty analysis of whether test scores rose.

Supervising or conducting evaluations requires an understanding of basic concepts and central issues in testing; such an understanding helps ensure that tests will not be misused as an overly simplistic "bottom line." When tests are chosen carefully and interpreted appropriately in evaluations, test results can help answer the question, "Is this project making a difference?"

Because testing is a complex technology, it is easy for program sponsors to assign concerns about how tests work or how they are constructed to the experts. But just as a wise patient would never undergo surgery without asking questions, those who intend to make use of test results need to pose pertinent questions about the costs, alternatives, and consequences of the testing decisions made in evaluations. This chapter will introduce some of the aspects of test use that ought to be looked into by those who commission and employ program evaluations.

The chapter opens with an explanation of what a test is. Subsequently, two types of testing will be considered with respect to evaluation. Traditional forms of testing, such as the multiple choice test, will be discussed first. Other forms of testing, the "alternative" forms of assessment which have received much recent attention, will also be discussed. The chapter will close with a list of questions to ask and issues to consider when using tests or interpreting test results in an evaluation.

What Is a Test?

Despite extensive experience with tests, many Americans, including some who regularly administer and use tests, would be lost if asked by that famous extraterrestrial, ET, "What is a test? We don't have these things on my planet." ET's confusion could be cleared up with an explanation of four concepts central to the definition of a test:

- A test focuses on a particular *domain* of interest.
- A test is a *sample* of behavior, products, answers, or performance from that domain.

- A test permits the user to make *inferences* about the larger domain of interest, and then, to use those inferences in *describing, making decisions, or determining consequences* about the test taker.
- The degree to which the specific inferences, descriptions, decisions, or consequences are appropriate is called *validity*.

Test Domain

A test is designed to measure a particular body of knowledge, skills, abilities, or performances which are of interest to the test user. This area of interest is called the test *domain* or test *universe*. The first step in constructing a test is to define the domain, so one can readily decide whether a particular aspect of knowledge, or a particular skill, task, ability, or performance clearly falls within the domain.

A straightforward, albeit somewhat simplistic, way to think about the domain for an achievement test is as a textbook, or as part of a textbook. For example, if a test writer wanted to construct a fourth-grade mathematics operations test, she might conceive of the test domain as the operations chapters from a typical fourth-grade mathematics textbook. The test domain could then be divided into four sections, called sub-domains or facets, representing the basic operations of addition, subtraction, multiplication, and division. Any of these four sub-domains could be specified further. For example, we might limit the addition facet to problems involving three or fewer digits. The sub-domains could also each be divided into numerical computational problems and word problems. Once we are satisfied with the domain specification, a test can be constructed to assess either the entire domain of fourth-grade arithmetic, or some facet of it.

The arithmetic example above represents a comparatively simple content or achievement domain in education. Not all test domains can be so easily defined, let alone divided so cleanly into sub-domains or facets. Furthermore, of course, test domains are not limited to academic or curricular areas. A test domain might focus on job-related skills for a particular occupation, for example, or on one of a wide range of more abstract traits such as intelligence, motivation, honesty, teacher competence, musical aptitude, mathematics problem-solving ability, or psychopathic deviation. Defining the test domain for an abstract trait is necessarily a more thorny undertaking than specifying the content of a typical textbook.

The concept of the test domain is relevant to evaluations in two important ways. First, too often people fail to question whether the domain is the correct one for the uses to which the test will be put. For example, imagine a standardized third-grade science test being used to evaluate the success of an innovative hands-on science curriculum. The test domain for the standardized test might, for example, cover facts concerning seeds and plants, matter and changes, rock formation, machines, weather, ecology, the moon, and health. The hands-on curriculum might, instead, emphasize skill development in the areas of observing, classifying, measuring, predicting, making generalizations, hypothesizing, and hypothesis-testing. While

nothing is inherently wrong with the facts-related standardized test, it was drawn from a domain of "third-grade basic science" that differs dramatically from the vision of "third-grade science" reflected in the hands-on science curriculum; thus, the standardized test's appropriateness in evaluating a hands-on curriculum should be carefully examined. The question, "Does this test cover the domain I am really interested in?" is central to proper test use.

A second major issue with respect to test domains in evaluation is the connotative power of the name given to a domain, and hence to its related test. Names of tests, such as those that are designed to measure "intelligence" or "functional literacy," can carry powerful cultural and personal meanings. These associative meanings color the way people use, interpret, and understand test performance. Thus, even when the definition of the test domain is appropriate for a given evaluation purpose, the name of the test may shape how test results are interpreted by various evaluation audiences.

A domain's name, for example, may fail to convey the uncertainty or sometimes the incompleteness of our conceptualizations. For example, people often forget, or may never know, that a particular "intelligence" or "teacher competency" test might represent only a small, and sometimes relatively unimportant, facet of a larger domain. Taking the test name too literally may mean that a person's test performance acquires all the generalized semantic, affective, connotative, emotional, and metaphorical baggage associated with the name of the particular domain — be it "honesty," "intelligence," or "readiness" — the test supposedly represents. Naming a test also affects attitudes about test use, sometimes at a profound level. For example, people resist the use of an "intelligence test" to retain children in kindergarten. However, when the same sort of test is labeled a "readiness" test, the practice becomes defensible and ultimately acceptable (Cunningham, 1988).

When a tester builds tests to measure constructs like intelligence, the only hope for some semblance of shared meaning is through clear communication of the specific facets of the domain the test is supposed to reflect. But this is easier said than done! The users of tests often do not reference test performance to the test developers' carefully crafted domain definition. Instead, they interpret test performance in terms of the contexts, meanings, purposes, and cultural sensibilities that they associate with a test's name. Thus, when choosing and using tests, we must be sensitive to the potential for misinterpretation based on the name of a test.

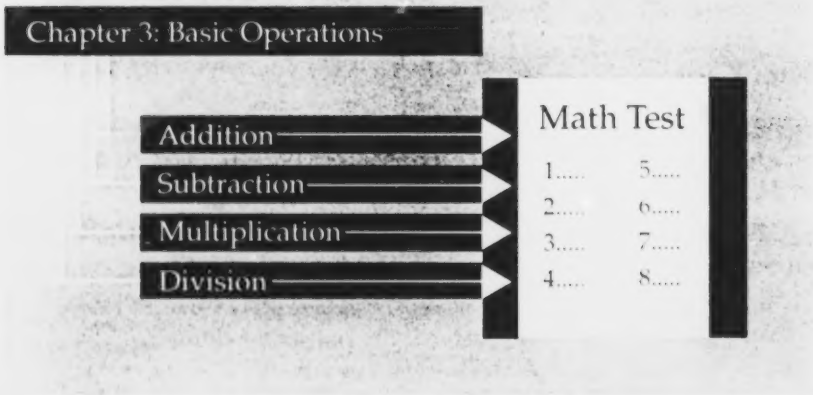
Sampling from the Test Domain

A second basic concept that needs to be explained when answering the question, "What is a test?" is that a test is a *sample* of behavior, products, answers, or performance from the larger domain of interest. Even for the comparatively simple domain of fourth-grade arithmetic problems, the number of possible test questions that could be constructed is staggering; we could never hope to ask students to solve all of them. Thus, we select a sample of problems to represent the important parts of the domain. It is this sample that constitutes the test of the domain.

Figure 3.1 illustrates the concept of sampling from the domain of fourth-grade arithmetic. The domain is represented by the chapter, with the four relevant sub-domains represented by chapter sub-headings. The test is made up of questions from the content of the chapter; in our illustration, each chapter part is represented by at least one question on the test. If an entire sub-domain, like "division," was not represented at all, or by only a few items, the representativeness of the sample would be called into question. "Does the sample of test questions adequately represent the domain?" is an important issue to address when using tests in evaluation.

Figure 3.1

Sampling from the Domain of Fourth-Grade Math Basic Operations



The sample of items that forms the test and is meant to represent the domain is generally developed according to plans, called test specifications. Test specifications describe in detail such matters as the type of items to be used, the number of items on the test, the proportion of test items representing each part of the domain, the time allocated to the test, and the statistical characteristics of the item such as item difficulty and readability levels. Test specifications, therefore, are the detailed blueprint for constructing the test. Well written test manuals typically include some of these details for the test user.

Making Inferences from Test Results

Implicit in the above discussion is the concept that it is the domain, not the test per se, which is of interest in any testing situation. Performance on the particular small sample of questions that constitute the test is of interest *only* in so far as it permits us to make *inferences* about the whole test domain. This concept of inferences is the third major component of the definition of a test: A test permits one to make *inferences* about the domain of interest, and then to use those inferences in describing, making decisions about, or determining consequences for the test taker, the institution, or the program.

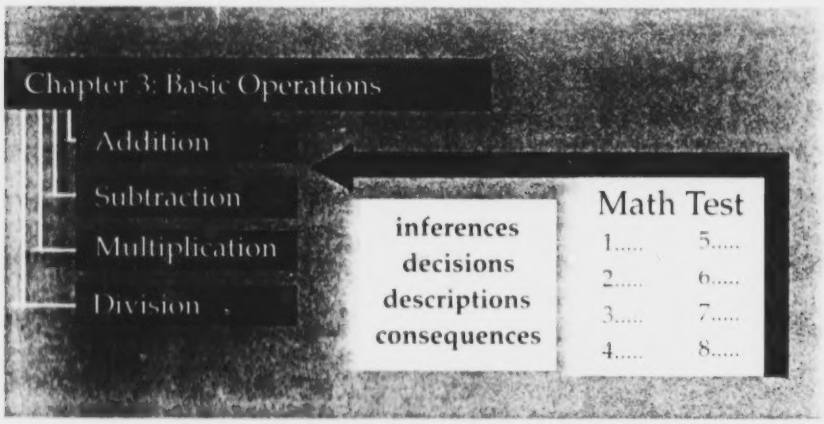
To continue with the example of the fourth-grade arithmetic test, a student's performance (or a class' average performance) on the ten, twenty, or one hundred problems making up the arithmetic test is never the ultimate concern. The ultimate concern is what the performance on the sample of problems, the test, suggests about the student's or the class' mastery of basic arithmetic (see figure 3.2).

Test results can lead to inferences that don't refer back to the original test domain but, instead, to a different domain. The content of aptitude tests, for example, is drawn from one domain, but the chief inference made from a student's score is

typically not about how fully he has mastered that domain; instead, the score is usually used to make an inference about the student's future performance. The SAT is a case in point: The items are drawn from mathematics and language arts domains, but the most common inference from the test scores — and the one for which the test was designed — concerns the likelihood of success in college. When SAT scores are used to determine eligibility for college athletics, to award scholarships, or to compare states on educational quality, the inferences are made to different domains altogether. The large black arrows in figure 3.3 represent inferences made to different domains from the SAT.

Figure 3.2

Making Inferences Back to the Domain of Fourth-Grade Basic Operations

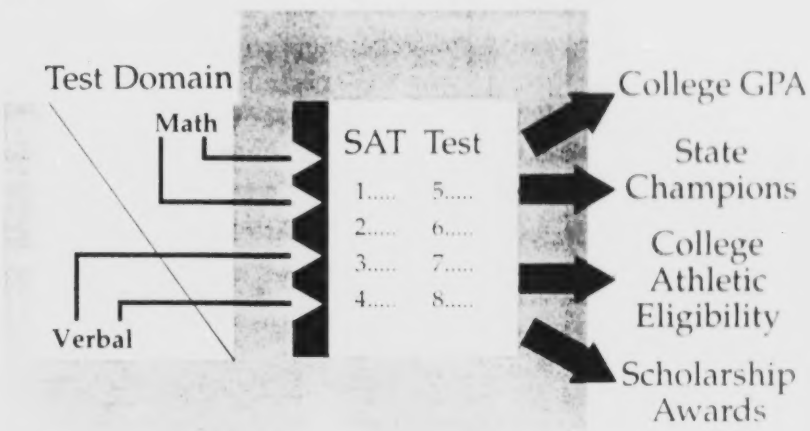


Test Validity

The final concept necessary to understanding a test is *validity*. Validity is the appropriateness, correctness, or meaningfulness of the specific inferences, descriptions, decisions, or consequences that are triggered by a test score. When the students in an innovative mathematics program get an average of 20 percent of the problems on the arithmetic test correct, and the project evaluator infers that they don't know very much about basic mathematics operations, the validity question is, "Is it correct to infer that the students haven't mastered arithmetic operations?" When the students are then all assigned to the remedial mathematics class for the remainder of the school year, the validity question is, "Is it appropriate to assign all students to remedial math (and hence, potentially, label them as low achievers) based on the program's average test score?"

Figure 3.3

Making Inferences from the SAT



Validity is a widely misused term. Some of the most common misunderstandings are clarified below.

1. *There is no such thing as a generically valid test.* In other words, it is incorrect to broadly and simply assert, "This is a valid test." Obviously, an inappropriate inference could be made from even the most well constructed and carefully administered test. Thus, statements about a test's validity must be qualified in

BEST COPY AVAILABLE

terms of the correctness of a *particular* inference and consequent description or decision about *particular* populations of test takers.

- *Validation is an ongoing, indeed, never-ending process* of accumulating evidence about validity. The fundamental characteristic of validation is the search for the *meaning* behind the test score. A true validation study seeks evidence that not only confirms, but also evidence that might cast doubt on the ability of the test to measure what it purports to measure.
- *The validity of a test is a matter of degree*, not a simple dichotomy of "valid" or "not valid." There no such thing as a perfect test; inferences are always problematic. Validation offers a reasoned defense for an inference, decision, or description, not proof.

Using Multiple Choice Tests in Evaluation

When constructing a test, a test writer must decide how the test takers will demonstrate what they know and can do. She can ask the examinees to select an answer from among several alternatives, as on a multiple choice test. Or, she could ask them to produce a response, as in answer to an essay question, and then evaluate the resultant product. She could also consider requiring that the students perform something, then assess the performance or observed process as it happens.

When using tests in evaluation, we can choose among these different modes of testing in order to answer our evaluation questions. Our choice of test type must be based on a clear understanding of the advantages and disadvantages of each mode, an awareness of the logistical and contextual constraints of the evaluation, and an understanding of the evaluation issues we want the tests to inform.

Asking examinees to select among alternatives, as in multiple choice or true/false examinations, has been the predominant mode of school testing in the United States for over four decades. Multiple choice tests are also the most common way of gathering achievement data in formal educational program evaluations — indeed, multiple choice standardized tests are mandated for evaluations of many federally-funded projects such as Chapter 1.

Multiple choice tests rose to their position of prominence for reasons that are easy to understand in historical context. During the first half of the 1800s in the United States, students graduating from high school typically took oral examinations. In Boston, educator Horace Mann found these examinations too time consuming. Further, the questions presented to each examinee had to be different; once an examinee exited the examination room, the questions presented to him could be revealed to the next examinees. Mann replaced the oral exams with essay examinations, which allowed the same questions to be administered to many students at once. The practice of giving essay examinations grew, and by the Civil War were the most common testing methodology in American schools.

Large scale immigration and greater access to schooling in the last decades of the nineteenth century dramatically increased the size of the school-going popula-

tion. Even more efficient means of testing were needed. What's more, studies had revealed that scores on essay tests varied greatly depending on who scored them. The invention of the multiple choice item, credited to Frederick Kelley in 1914, solved the problems of inefficiency and subjectivity in essay testing. The use of multiple choice tests grew rapidly, especially after the development of optical scanners in the 1950s. Efficiency and objectivity still head the list of advantages of multiple choice tests.

Advantages of Using Multiple Choice Tests in Evaluation

Multiple choice tests are recommended for use in evaluation for many reasons. Some of the most common are listed below.

- They are objectively scored: no matter which person, or for that matter, which machine corrects the test, the scores will be the same.
- In addition, multiple choice tests are extremely efficient; they can be administered to many students at a time and can be scored quickly, accurately, and inexpensively thanks to optical scanners. There is typically no need for trained administrators or subject-area specialists.
- Multiple choice tests can cover a great deal of the test domain in comparatively little time. For example, students can answer fifty questions in forty minutes.
- Multiple choice tests often possess a desirable test characteristic called reliability. Reliability refers to the consistency of test scores across different testing conditions or different forms of a test (see box below).

A Closer Look

Reliability

Reliability refers to the accuracy, consistency, or lack of measurement error associated with a set of test scores. If you stepped on your bathroom scale at 6:00 a.m. and got on it again at 7:00 a.m. the same morning and got two very different readings, you would question its reliability. Similarly, the trustworthiness of a test would be called into question if it yielded very different results under slightly different testing conditions. Of course no test, or bathroom scale, or opinion poll, or measuring instrument of any sort is without some error; you just want to be sure that the accuracy is adequate for the decisions you want to make from test results.

Perfect reliability is 1.00; complete lack of reliability is 0.00. Many tests used for important decisions at the individual level, such as the SAT, have reliabilities in the low .90s. One rule of thumb given for tests that are used for decisions about groups, as is usually the case in evaluations, is that they should have reliabilities above .65 (Mehrens & Lehmann, 1987).

- Multiple choice tests, in part because they are objectively scored and widely used, are perceived by many evaluation audiences to be good, credible sources of information.
- Multiple choice tests provide scores in metrics that are familiar to many evalua-

tion audiences. For example, percentiles or grade-equivalent scores are two norm-referenced metrics that are widely reported in evaluations (see below).

A Closer Look

Norm-Referenced versus Criterion-Referenced Scores

Norm-referencing and criterion-referencing are two different ways to report test scores. Nothing is inherently better about either method of attaching meaning to test performance; they simply serve different purposes.

Norm-referenced scores compare an examinee's test performance to the performance of other examinees. In routine classroom testing, a student's score would typically be compared to those of her classmates: "Maria performs above the class average in mathematics, but is in the lower third in reading." In standardized educational testing or employment testing, an examinee's performance is compared to the performance of a clearly defined reference group of examinees called a norming group. The scores of the norming group are used to devise test norms — that is, the data about normal, below normal, and above normal performance.

Test users need to know that the norms for a test are appropriate. For example, if the norms were developed over a decade ago, the test user needs to question whether comparing a person's score with that group is reasonable. Norm groups should also be representative; for example, national norms should be derived from a group of examinees who are representative of the United States population with respect to such characteristics as gender, age, race, and region of the country.

Criterion-referenced scores, on the other hand, say something about how the examinee performed relative to an absolute performance standard or a criterion. Ideally, a criterion-referenced score provides a clear answer to the question, "What does this student know?" or "What can this student do?" For example, if a content domain is thoroughly defined, and if the test items sampled from the domain are representative, then a student's score of 80 percent correct should be interpretable in terms of how much of the content domain the student knows. Other criterion-referenced interpretations might indicate what percent or how many of the course objectives were passed. For some professionals, criterion-referencing is synonymous with mastery testing, in which individual test performance is described as "pass" or "fail" relative to a pre-determined cut-off score. Interestingly, such a cut-off score is sometimes norm-referenced at its root, in that it is determined by considering the performance of a norming group.

Many standardized achievement batteries provide both norm-referenced and criterion-referenced interpretations of student performance. For example, a student's performance on the language mechanics portion of a test could be expressed as a national percentile, which compares his performance to a national norming group. It could also be expressed as a local percentile, which indicates how well he did compared to the other students in the district. At the same time, the score report might provide some criterion-referenced interpretations: This student demonstrated mastery of sentence-final punctuation, partial mastery of pronoun usage, and non-mastery of capitalization.

Example from classroom testing: A 50-item mathematics operations test is given to a class of 25 fourth-grade students. In this class, Rosalyn got 42 questions right. Twenty of Rosalyn's classmates scored below 42.

Criterion-referenced scoring of Rosalyn's performance:	*Norm-referenced scoring of Rosalyn's performance:*
Rosalyn got 43 out of 50 questions right.	Rosalyn scored at the 80th percentile.
Rosalyn got 84%.	Rosalyn ranked 6th in her class.

Disadvantages of Using Multiple Choice Tests in Evaluation

Despite their many advantages, multiple choice tests have been criticized for some of the following reasons.

- Multiple choice test items are often inherently ambiguous. Wording that may appear clear to a test writer may unintentionally confuse the test taker.
- Multiple choice tests provide little truly diagnostic information about students. They provide no information about *why* pupils get items right or wrong and no data about the process that the students employed in responding.
- In a similar sense, typical multiple choice standardized tests provide little information that can be directly used by the teacher to guide or improve instruction.
- Multiple choice tests often do not tap (hence do not provide information about) students' higher order thinking skills. It is easier to write multiple choice items to measure factual knowledge than complex, multi-step mental processes, though the latter can be done.
- If multiple choice tests are associated with important sanctions and rewards, such as promotion to the next grade or graduation from high school, they can exert negative influences on the curriculum. Imagine a school that wishes to receive continued project funding for an innovative program they've developed. If continued funding is contingent on test score improvement, we would almost certainly see such an improvement. Unfortunately, the reason for a score increase under such circumstances is often that instruction has been reduced to drilling for the test.

Using Alternative Assessments in Evaluation

The well documented problems with multiple choice tests, such as those highlighted above, have provoked a demand for different kinds of tests. These "different" kinds of tests are generally referred to as "alternative assessments," since they represent an alternative to the dominant multiple choice standardized tests. Performance assessments, portfolios, and even multiple choice items that require higher-order thinking have all been included under this rubric.

What these alternative test and item types have in common are the following characteristics:

- They focus on a whole task, not on discrete bits of information.
- They require that the examinee produce or perform, not select.
- They are designed to provide information that both the learner and the teacher can use to improve the teaching/learning process.

- The testing tasks are designed to be direct, realistic reflections of the real-world situations learners are likely to encounter.
- The testing tasks can be integrated into the instructional process — in fact, they are often indistinguishable from the kinds of things teachers do in classrooms as part of regular instruction.

The forms that such alternative or, as they are sometimes called, "authentic" assessments can take are quite varied; two examples are provided in the box below. It should be clear from these examples that alternative assessments are not necessarily new. Oral exams, essay tests, and hands-on science laboratory work, all time-honored educational practices, could easily be termed alternative or authentic assessments.

A Closer Look

Examples of Alternative Assessment

Performance assessment requires students to demonstrate their knowledge of particular subject matter. These types of assessment are commonly encountered in science in the form of experiments and other laboratory work, but they can be used in any subject area. For example, in social studies, a unit on reading and using a compass could culminate in an outdoor orienteering contest to find a "buried treasure." Given a list of directions and a compass, a student would have to follow the directions successfully from start to finish to locate the treasure.

Portfolios are simply repositories for student work. Ideally, they can showcase and document a student's development in a given subject over time. Typically, the overall format is decided upon by the teacher, but the student has at least some role in selecting which work to include and determining how that work will be evaluated. For example, a writing portfolio might include several drafts of a several different papers, as well as an essay in which the student explicitly compares two of the papers. Later in the semester, the student might write a table of contents for the portfolio and an evaluative essay about the contents and the development they reflect.

Advantages of Using Alternative Assessments in Evaluation

Because they are defined essentially as "not multiple choice," alternative assessments present distinct advantages in evaluation over traditional standardized tests.

- Alternative assessments can provide rich data about students' thinking processes.
- Data from alternative assessments can guide instruction and provide information useful in formative evaluations.
- Alternative assessments can be powerful tools in the hands of good teachers, if given by teachers as part of regular instruction.

- Alternative assessments employ and engage a valuable classroom resource — the teacher, who has an unparalleled knowledge of classroom context. Further, involving teachers can encourage their cooperation in the program and the evaluation.
- Alternative assessments typically target higher-order skills.
- If it is true that measurement shapes instruction, then alternative assessments hold the promise of reorienting instruction in a desired direction.
- Alternative assessments have face validity for students: Because such assessments (ideally) resemble and are integrated with instruction, they look to students like reasonable tests ought to look.

Disadvantages of Using Alternative Assessments in Evaluation

While alternative assessments are valuable classroom tools, they present major problems when used in evaluations or in policy decisions on a large scale. Some of the problems are discussed below.

- The logistics of administering individual oral exams, observing individual or small group performances, or setting up complex science experiment stations are complicated. For example, what does a teacher do with the other twenty or more students in a class when he is testing children individually for fifteen minutes each?
- It is difficult to standardize the administration of alternative assessments. For example, results may vary depending on who is administering the test (the teacher or an outside person), where it is administered (the classroom, a library, the hallway, or an empty room), and with what equipment (differing laboratory set-ups). The lack of standardization renders results not comparable from one setting to the next.
- Alternative assessments are comparatively inefficient and expensive; they usually require trained administrators and/or scorers and they take a great deal of time to give and score.
- Typically, alternative assessments do not sample broadly from the content domain. For example, a student who fills her math portfolio with examples of her "best work" might by design or oversight not include anything that would serve as evidence of her ability to multiply and divide.
- The issue of differential test performance by various socioeconomic or racial/ethnic groups does not disappear with the advent of alternative assessment. On the contrary, the achievement gap between various groups on alternative tests has been found to be greater than on traditional tests (Nuttall & Goldstein, 1990, cited in Maeroff, 1991).

Questions to Ask When Using Tests in Program Evaluation

We have to use tests and assessment instruments, but we must not deceive ourselves or others into believing that our test results are infallible. Wise test use involves asking informed questions about every testing decision. None of the disadvantages mentioned above with respect to either traditional or alternative testing presents insurmountable difficulties as long as we know about the potential pitfalls and know to ask about them.

When making testing decisions in evaluation, or working with an evaluator who is making testing decisions, some of the questions you might ask are:

- What are our purposes for testing? What inferences, decisions, descriptions, or consequences do we hope to produce from test results?
- Is the domain of the test the domain we're interested in? Does the name of the test mask an incomplete conception of the domain?
- Is the norm group for the test appropriate for our testing decisions? Are the norms up-to-date?
- Are we using an appropriate metric and ensuring that those who use our evaluation results understand it?
- Are there other indicators that can support the conclusions drawn from the test results? What other sources of information (such as other tests or teacher opinions) can help us judge the inferences we make from the test?
- Do we have to test every student in every grade to make our decisions? If you're using an expensive or disruptive testing procedure, and you don't need to make decisions at the level of the individual student (that is, "Should Johnny do more work on addition?"), consider testing only a random sample of the students (see boxed insert opposite).
- How can we tap the context-rich knowledge of the teacher?
- What contextual and logistical constraints will affect our choice of testing mode?

A Closer Look

Matrix Sampling

"Do the fifth graders in our innovative reading program now read better?" This is a common sort of question to ask in an evaluation; the most common way to answer it has been to choose a test, administer it to all fifth graders in the program, and look at the results.

But think about this: When a banana boat arrives in the harbor from a foreign country, does the agricultural inspector open every banana in order to determine if the shipment is up to standards? He or she samples from among the bananas in the boat, and makes inferences about the whole shipment based on the bananas in the sample. We can apply the same principle in educational evaluation. We don't have to test everyone in a group in order to make inferences about that group. If we are using an expensive, time consuming, or disruptive testing procedure, sampling becomes an especially attractive option.

Another variation on sampling is called matrix sampling. In matrix sampling, both students and types of assessment are systematically sampled. For example, of the one hundred fifth graders in our innovative reading program, a randomly selected twenty-five might take a multiple choice test of reading skills, another randomly selected twenty-five might provide oral reading samples, twenty-five might retell a story they'd read, and the last twenty-five might turn in portfolios of book reports. Matrix sampling allows us to use a wide range of assessment techniques and thus increases the scope of information we have about the targeted skill or subject area, in this case, reading.

Conclusion: Learning from Funded Projects

Evaluation is a complex and powerful tool. As we have seen, it embodies deeply held beliefs about the object of evaluation and, in turn, helps shape those beliefs in ways both intended and unintended. We have also seen that there is no single right approach to educational program evaluation (or to educational testing, for that matter). This means that every evaluation design involves choices and decisions, explicit or implicit. And if the politics of designing an evaluation are challenging, the politics of reporting on one can be positively daunting. Furthermore, for contributions officers faced with limited budgets, evaluation must be considered simultaneously as an additional expense and a potential economy. Yet despite these complexities, only systematic, appropriate evaluations will tell corporate funders, over time, what kind of progress they're making — toward improved student learning, toward the advancement of long-range education reforms, and toward the achievement of corporate philanthropic and community relations goals.

So, like any powerful tool, evaluation comes with cautions and guidelines for best use, some of which we've tried to cover in this report:

- Recognize the assumptions and values you and your evaluator bring to bear when posing and answering evaluation questions. Evaluation is not a science; its course and conclusions are greatly affected by the way we consciously or unconsciously think about it.
- Become familiar with existing evaluation approaches, and remain flexible in considering those components that are most appropriate for the context and constraints of your evaluation.
- Stay actively involved in the evaluation, both directly and by knowing what questions to ask of yourself and your evaluator.

Undertaken thoughtfully and carefully, educational program evaluation represents an opportunity to understand and enhance the role of corporations in education reform. For your individual company, it represents an opportunity to improve project and program performance. And it signals, to company staff and to the company's various publics, that your commitment to education includes a willingness to reflect on your own activities and to modify them if necessary to produce better results.

A Closer Look

Contributing to the Knowledge Base

More can be learned from our evaluations when they are seen, not as isolated endeavors, but as part of an ongoing inquiry into educational programs and practices. The results of any one evaluation often take on additional meaning when viewed in terms of other evaluations of other educational endeavors. Thus, anyone undertaking a serious evaluation of an educational initiative should be encouraged to share results and reports not just with the audiences who have a direct stake in the program, but also with the broader audience of professional and public communities interested in educational research and improvement. While there are innumerable avenues for sharing evaluation results, here we mention just three.

First, the four most common vehicles for communication with the public are the press, parent/teacher organizations, school boards, and education advocacy groups. Four umbrella organizations that can provide you with leads for disseminating your results to the public are: the Education Writers Association, the National Parent Teacher Association, the National School Boards Association, and the National Coalition of Advocates for Students.

Second, there are literally dozens of professional associations concerned with education improvement. Some of these focus on subject matter areas, such as the National Council of Teachers of Mathematics, and some focus on particular professional roles, such as the American Federation of Teachers, the National Education Association, or the Association for Supervision and Curriculum Development. The professional organization that is probably the preeminent vehicle for communication of educational research and evaluation results is the American Educational Research Association (AERA), one of the sponsors of the evaluation *Standards*. AERA and its numerous special interest groups publish literally dozens of journals and newsletters spanning the full range of educational levels and topics.

Third, and easily the most prominent repository of information of educational research and evaluation, is the Educational Resources and Information Clearinghouse (ERIC). ERIC's database, the world's largest source of education information, contains citations of more than 700,000 documents and journal articles on education research and practice. The ERIC database is available to the public at more than 3,000 locations worldwide. Practi-tioners, librarians, policymakers, researchers, and students can access information on a wide range of subjects, such as "academic excellence," "merit pay," "school security," and "educational innovations."

A new component of the ERIC system called ACCESS ERIC provides comprehensive information on using ERIC. Available toll free (1-800-USE ERIC), ACCESS ERIC can keep you informed of the wealth of information offered by ERIC and other education information service providers. ACCESS ERIC staff will answer questions, refer callers to subject-specific information sources, and provide ERIC publications. ACCESS ERIC can also tell you how to submit your evaluation report for possible inclusion in the ERIC system. So if you are seriously interested in learning from evaluation of educational programs, not just for yourself or your corporation, but for the broader community of citizens and professionals interested in education, we urge you not only to use ERIC but to submit your evaluation studies to it, so that others might learn from your efforts.

References

Alkin, M. C., & Stecher, B. (1983). A study of evaluation costs. In Alkin, M. C., & Solmon, L. C. (Ed.), *The costs of evaluation*. (pp. 119-132). Beverly Hills, CA: Sage Publications.

Brinkerhoff, R. O. (1983). *Program evaluation: A practitioner's guide for trainers and educators*. Boston: Kluwer-Nijhoff.

Cooley, W. W., & Bickell, W. E. (1986). *Decision-oriented educational research*. Boston: Kluwer-Nijhoff.

Denny, T. (1978). Story telling and educational understanding (occasional paper #12). Evaluation Center, Western Michigan University, Kalamazoo, MI.

Doyle, D. P. (1991, September). Empowering teachers. *The Atlantic Monthly*, p. 15.

Eisner, E. W. (1983). Educational connoisseurship and criticism: Their form and functions in educational evaluation. In Madaus, G. F., Stufflebeam, D. L., & Scriven, M. S. (Ed.), *Evaluation models: Viewpoints on educational and human sevices evaluation*. (pp. 335-348). Boston: Kluwer-Nijhoff.

Eisner, E. (1985). *The art of educational evaluation: A personal view*. London: The Falmer Press.

Eisner, E. W. (1991). *The enlightened eye*. New York: Macmillan.

Ellwein, M. C., & Graue, M. E. (in press). Knowing children. In Grant, C., & Gomez, M. L. (Ed.), *Campus and classroom: Making education multicultural*. New York: Macmillan.

Fink, A., & Kosecoff, J. (June 1983). How to evaluate on a shoestring. *How to evaluate education programs, Chapter 70*, pp. 1-8.

Guba, E., & Lincoln, Y. (1981). *Effective evaluation*. San Francisco: Jossey-Bass Publishers.

Guba, E., & Linclon, Y. (1983). Epistomological and methodological bases of naturalistic inquiry. In Madaus, G. F., Stufflebeam, D. L., & Scriven, M. S. (Ed.), *Evaluation models: Viewpoints on educational and human sevices evaluation*. (pp. 311-334). Boston: Kluwer-Nijhoff.

Herman, J. L., Morris, L. L., & Fitz-Gibbon, C. T. (1987). *Evaluator's handbook*. Newbury Park, CA: Sage Publications.

Jackson, P. W. (1968). *Life in classrooms*. New York: Holt, Rinehart and Winston.

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials*. New York: McGraw Hill.

Kidder, T. (1989). *Among schoolchildren*. New York: Avon.

Kliebard, H. M. (1972). Metaphorical roots of curriculum design. *Teachers College Record*, 73(3), pp. 403-404.

Kritzman, L. D. (Ed.). (1990). *Michel Foucault: Politics, philosophy, culture: Interviews and other writings* 1977 – 1984. New York: Routledge.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Madaus, G. F., Stufflebeam, D. L., & Scriven, M. S. (Ed.). (1983). *Evaluation models: Viewpoints on educational and human services evaluation*. Boston: Kluwer-Nijhoff.

Madaus, G. F., & Stufflebeam, D. L. (1989). *Educational evaluation: Classic works of Ralph W. Tyler*. Boston: Kluwer Academic.

Maeroff, G. I. (1991). Assessing alternative assessment. *Phi Delta Kappan*, 73(4), pp. 272-281.

Mehrens, W. A., & Lehmann, I. J. (1987). *Using standardized tests in education*. (4th ed.). New York: Longman.

Nitko, A. J. (1983). *Educational tests and measurement: An introduction*. New York: Harcourt Brace Jovanovich, Inc.

Nuttall, D. L., & Goldstein, H. (1990). The 1988 examination results for ILEA. Paper presented to the Inner London Educational Authority Committees.

Provus, M. N. (1971). *Discrepancy evaluation*. Berkeley, CA: McCutcheon.

Ralph, J., & Dwyer, C. M. (1988). *Making the case: Evidence of program effectiveness in schools and classrooms - Criteria and guidelines for the US Department of Education's Program Effectiveness Panel*. Washington, DC: US Government Printing Office.

Sanders, J. R. (1983). Cost implications of the standards. In Alkin, M. C., & Solmon, L. C. (Ed.), *The costs of evaluation*. (pp. 101-118). Beverly Hills, CA: Sage Publications.

Sarason, S. (1990). *The predictable failure of educational reform*. San Francisco: Jossey Bass.

Scriven, M. (1981). *Evaluation thesaurus*. (3rd ed.). Inverness, CA: Edgepress.

Sizer, T. (1992). *Horace's school*. Boston: Houghton Mifflin.

Smith, E. R., & Tyler, R. W. (1942). *Appraising and recording student progress*. New York: Harper and Row.

Smith, M. R. (1987). Army ordnance and the 'American System' of manufacturing: 1815–1861. In Smith, M.R. (Ed.), *Military enterprise and technological changes*. (pp. 39-86). Cambridge, MA: MIT Press.

Stake, R. (1983a). Program evaluation, particularly responsive evaluation. In Madaus, G. F., Stufflebeam, D. L., & Scriven, M. S. (Ed.), *Evaluation models: Viewpoints on educational and human services evaluation*. (pp. 287-310). Boston: Kluwer-Nijhoff.

Stake, R. (1983b). The case study method in social inquiry. In Madaus, G. F., Stufflebeam, D. L., & Scriven, M. S. (Ed.), *Evaluation models: Viewpoints on educational and human services evaluation*. (pp. 279-286). Boston: Kluwer-Nijhoff.

Staudenmaier, J. M. (1985). *Technology's storytellers: Reweaving the human fabric*. Cambridge, MA: MIT Press.

Staudenmaier, J. (1988). *Technology and faith* (audio cassette). Kansas City, MO: Credence Cassettes.

Staudenmaier, J. M. (1989). U.S. technological style and the atrophy of civic commitment. In Gilpi, D. L. (Ed.), *Beyond individualism toward a retrieval of moral discourse in America*. (pp. 120-152). South Bend, IN: Notre Dame Press.

Stufflebeam, D. L., & Madaus, G. F. (1983). The standards for evaluation of educational programs, projects and materials: A description and summary. In Madaus, G. F., Scriven, M. S., & Stufflebeam, D. L. (Ed.), *Evaluation models: Viewpoints on educational and human services evaluation*. (pp. 395-404). Boston: Kluwer-Nijhoff.

Stufflebeam, D. L., & Shinkfield, A. J. (1985). *Systematic evaluation*. Boston: Kluwer-Nijhoff.

Thomas, L. (1980). *The medusa and the snail: More notes of a biology watcher*. New York: Bantam Books.

Tyack, D. B. (1974). *The one best system: A history of American urban education*. Cambridge, MA: Harvard University Press.

Worthen, B., & Sanders, J. R. (1987). *Educational evaluation: Alternative approaches and practical guidelines*. New York: Longman.

Glossary

Accountability: Setting correct goals, evaluating their degree of achievement at what price, presenting and interpreting this information to the public, and accepting responsibility for any results that are perceived to be inadequate (Mehrens & Lehmann, 1987).

Achievement test: A test that measures formal school learning.

Alternative (authentic) assessments: Tests that represent an alternative to the dominant multiple choice standardized tests. Examples include performance assessments, portfolios, and multiple choice tests that require higher-order thinking.

Aptitude test: A test that measures ability to learn new tasks, or future performance.

Assessment:* The act of determining the standing of an object on some variable of interest. For example, testing students and reporting raw scores.

Audiences:* Those persons who will be guided by the evaluation in making decisions and all others who have a stake in the evaluation.

Bias:* In evaluation, a consistent alignment with one point of view.

Black box: A shortcoming of the goals-oriented evaluation approach: the failure of this approach to look at resources associated with the program or at the educational processes themselves.

Case study:* An intensive, detailed description and analysis of a single project, program, or instructional material in the context of its environment.

Client:* The individual, group, or organization that hires the evaluator.

Context (of an evaluation):* The combination of the factors accompanying the study that may have influenced its results. These factors include the geographic location of the study, its timing, the political and social climate in the region at that time, the other relevant professional activities that were in progress, and any existing pertinent economic conditions.

Contract:* A written or oral agreement between the evaluator and the client which is enforceable by law. It is a mutual understanding of expectations and responsibilities for both parties.

Cost-free evaluations: Funded evaluations should be able to show a net positive return over time. They should, on average, more than pay for themselves in terms of program improvement.

Criterion-referenced tests:* Tests whose scores are interpreted by referral to specifically defined performances, rather than by referral to the performance of some comparable group of people.

Decision-oriented approach: An evaluation approach that delineates, obtains, and applies descriptive and judgmental information concerning a program's merit as revealed by its goals, design, implementation, and results for purposes of decision making and accountability. Examples include IPO and CIPP models.

NOTE: An asterisk (*) indicates that the definition was taken from the *Standards* (Joint Committee, 1981).

Diagnostic test:* A test primarily concerned with determining problems or deficiencies of the test taker in a specific range of instructional content in order to prescribe placement or remediation.

Dissemination:* The communication of the actions — by written, oral, and/or audio-visual reporting — of evaluators to foster knowledge of the evaluation findings among all right-to-know audiences.

Educational objectives: Clear statements of what it is we want students to know or be able to do as a result of a program.

Eight-Year Study: A study directed by Ralph Tyler in the 1930s, funded by the Carnegie Corporation. It was the only large study of the differential effectiveness of various types of schooling until well after World War II. It introduced educators throughout America to a new and broader view of educational evaluation.

Evaluation:* Systematic investigation of the worth or merit of an object (a program, project, or instructional material).

Evaluator: Anyone who accepts and executes responsibility for planning, conducting, and reporting evaluations.

Formative evaluation:* Evaluation designed and used to improve an object, especially when it is still being developed.

Front-end analysis: The research (interviews, document review, etc.) that guides the evaluation design. It should always involve identification of the political context of the project.

Goals-oriented approach: An evaluation approach that considers only predetermined program goals as objects of an evaluation.

Grade-equivalent scores: Scores that reflect the grade placement for which a raw score on a test is average (Nitko, 1983).

Higher-order thinking skills: Skills that require the learner to produce or create something, rather than simply to answer questions. These skills are usually described in terms of the highest three categories of Bloom's Taxonomy — analysis, synthesis, and evaluation (Nitko, 1983).

Information needs:* Information requirements of the clients and other pertinent audiences to be met by the evaluation.

IPO Model: A decision-oriented evaluation model that is concerned with the inputs, processes, and outcomes (or product) of educational programs.

Linear (standard) scores: Transformed scores for which the resulting set of values has a distribution shape identical to the original raw-score distribution (z- and t-scores are examples).

Matrix sampling: A sampling technique that systematically samples both students and types of assessment, allowing use of a wide range of assessment techniques with minimal disruption.

Measurement: A procedure for assigning numbers (scores) to a specified attribute or characteristic of persons in such a way as to maintain the real-world relationships among the persons with regard to the attribute being measured (Lord & Novick, cited in Nitko, 1983).

Merit:* The excellence of an object as assessed by its intrinsic qualities or performance.

Monitoring: Overseeing the progress of instructional activites, deciding if these activities are effective or whether different activites are warranted (Nitko, 1983).

National Diffusion Network (NDN): a mechanism funded by the federal government as a way to disseminate exemplary programs and products.

Needs assessment: An evaluation that compares the "is" of a system with the "should be."

NOTE: An asterisk (*) indicates that the definition was taken from the *Standards* (Joint Committee, 1981).

Norm group: A clearly defined group to which test scores are compared.

Norm-referenced scores: Scores that compare an examinee's test performance to the performance of other examinees.

Object (of an evaluation):* What one is evaluating (a program, project, instructional materials, etc.).

Percentile: A point on the distribution below which a certain percentage of the scores fall.

Percentile rank: A score that gives a person's relative position or the percentage of students' scores falling below that individual's obtained score.

Performance assessment: An alternative assessment that requires students to demonstrate their knowledge of particular subject matter.

Portfolios: A repository for student work; a type of alternative assessment.

Preordinate approaches: Evaluations that emphasize statement of goals, use of objective tests, standards held by program personnel, and research-type reports.

Program Effectiveness Project (PEP): The project (and panel) that grew out of the joint dissemination review panel founded by the federal government to ensure that before spending federal funds on dissemination, there is credible evidence of program success.

Program evaluations:* Evaluations that assess educational activities which provide services on a continuing basis and often involve curricular offerings. Examples include evaluation of a school district's reading program, a state's special education program, and a university's continuing education program.

Progressive education movement: A renewal in education after the Great Depression which reflected the philosophy of pragmatism and employed tools from behavioristic psychology.

Purposes (of an evaluation):* The objectives of an evaluation (to judge the relative merits of competing textbooks, for example, or to monitor and report on how well a project plan is implemented) and the intended uses of its reports (to help teachers choose a textbook, perhaps, or to help a school district carry out a special project).

Qualitative information:* Facts and claims presented in narrative, not numerical, form.

Quantitative information:* Facts and claims that are represented by numbers.

Raw score: A person's observed score on a test, reported in total questions or points correct. A raw score on its own usually gives little information about a person's performance.

Reliability: The accuracy, consistency, or lack of measurement error associated with a set of test scores.

Responsive (or naturalistic) evaluation models: Evaluation models that orient more directly to program activities than to program intents, respond to audience requirements for information, and refer to the different value perspectives of the people at hand in reporting the success and failure of the program.

Standard: A principle commonly agreed to by experts in the conduct and use of evaluation for the measure of the value or quality of an evaluation.

Standards for Evaluations of Educational Programs, Projects, and Materials: Widely recognized guidelines that can be used in all phases of an evaluation. They are grouped according to four attributes — utility, feasibility, propriety, and accuracy.

Standardized test:* A sample of items or situations with definite directions for administration and scoring. Most often accompanied by data on reliability and validity and sometimes by normative information.

NOTE: An asterisk (*) indicates that the definition was taken from the *Standards* (Joint Committee, 1981).

Summative evaluation:* Evaluation designed to present conclusions about the merit or worth of an object and recommendations about whether it should be retained, altered, or eliminated.

Test: A sample of behaviors, products, answers, or performances from a specified domain of content that permits the user to make inferences about that domain of interest.

Test domain: A particular body of knowledge, skills, abilities, or performances which are of interest to a test user.

Tyler Triangle: Ralph Tyler's graphic portrayal of the evaluation process according to his model. The apex of the triangle represents the educational objectives, and the two legs represent the educational experiences and the determination of the attainment of the educational objectives.

Validity:* The soundness of the use and interpretation of a measure; the appropriateness, correctness, or meaningfulness of the specific inferences, descriptions, decisions, or consequences that are triggered by a test score.

Worth:* The value of an object in relationship to a purpose.

NOTE: An asterisk (*) indicates that the definition was taken from the *Standards* (Joint Committee, 1981).

Testing in Educational Program Evaluation

What is a test?

- A test focuses on a particular **domain** of interest.
- A test is a **sample** of behavior, products, answers, or performance from that domain.
- A test permits the user to make inferences about the larger domain of interest, and then to use those inferences in **describing or making decisions about** the test taker.
- The degree to which the specific inferences, descriptions, or decisions are appropriate is called **validity**.

Testing Modes

Traditional (multiple choice) tests:

ADVANTAGES	DISADVANTAGES
• objectively scored	• questions are sometimes ambiguous
• relatively nondisruptive	• provide little diagnostic information
• less expensive	• rarely tap higher-order skills
• efficient	• negative influence on curriculum if used for drilling
• familiar and credible to many audiences	

Alternative assessments (portfolios, performance assessment):

ADVANTAGES	DISADVANTAGES
• focus on a whole task	• relatively disruptive
• provide rich data about students' thinking processes	• expensive, less efficient
• designed to reflect real-world situations	• lack standardization
• can be integrated into the instructional process	• typically don't sample broadly from content domain
• provide information that can be used to improve the learning process	

(continued on reverse side)

Questions to ask when using tests in program evaluation:

- What are our reasons for testing? What inferences, decisions, or descriptions do we hope to derive from test results?
- What kinds of tests are most likely to produce results that could support those inferences and decisions?
- What contextual and logistical constraints will affect our choice of testing modes?
- Is the actual domain of the test the domain we're interested in? Or does the name of the test (e.g., intelligence test) mask an incomplete or imprecise conception of the domain?
- Is the norm group for the test appropriate for our testing decisions? Are the norms up to date?
- Are there other indicators that can support the conclusions drawn from the test results? What other sources of information (e.g., teacher opinions) can help us judge the validity of our inferences, and how can we tap those sources?
- Do we have to test every student in every grade to make our decisions, or should we consider testing a random sample?

NOTE: The ideas and information presented here are excerpted from a recent report, *Testing and Evaluation: Learning from the Projects We Fund*, by George Madaus, Walt Haney, and Amelia Kreitzer. The full report is available for $15 prepaid, from the Council for Aid to Education, 51 Madison Avenue, Suite 2200, New York, New York 10010.

COUNCIL FOR AID TO
EDUCATION

Playing an Active Role in Evaluation Tasks

Every evaluation may be seen as involving three phases and ten tasks. As a contributions administrator or program officer, you'll be involved primarily in the planning and use of the evaluation (phases 1 and 3).

PHASE 1. PLANNING

- deciding whether or not to do an evaluation
- clarifying and assessing the purposes of the evaluation
- ensuring the political viability of the evaluation study

PHASE 2. MANAGING

- drawing up the contract
- staffing the study
- managing the study
- collecting data
- analyzing data

PHASE 3. USING

- reporting the findings
- applying results

Among the important questions you might ask in planning and using evaluations:

- Why evaluate? What can we expect to learn? Does the evaluation design maximize our chances of learning what we need to know?
- Who are the relevant audiences? Does the evaluation design speak to the needs of multiple audiences?
- Will the evaluator be credible to relevant audiences?
- Does the design allow for collection of data from different sources in different ways?
- Do the evaluation plan and the budget both include provisions for reporting to different audiences, in ways appropriate to each?
- How can we include the results of the evaluation in relevant bodies of knowledge and collections of research?

NOTE: The ideas and information presented here are excerpted from a recent report, *Testing and Evaluation: Learning from the Projects We Fund*, by George Madaus, Walt Haney, and Amelia Kreitzer. The full report is available for $15 prepaid, from the Council for Aid to Education, 51 Madison Avenue, Suite 2200, New York, New York 10010.

COUNCIL FOR AID TO EDUCATION

Three Approaches to Educational Program Evaluation

Begin by considering the evaluation questions you'd like to ask, the issues and audiences that must be addressed, and the resources available for the evaluation. Then combine elements from different evaluation models to provide the evidence that would be most appropriate for your purposes and to meet the information needs of as many stakeholders as possible, given budgetary and other constraints.

Goals-Oriented Evaluation

Assumes that the chief purpose of evaluation is to determine whether or not the program's stated goals or objectives have been achieved — does the quality of the "product" measure up to predetermined standards? Typically employs standardized tests.

ADVANTAGES

- widely used and credible
- encourages clarity about goals, as well as accountability
- minimizes disruption
- relatively inexpensive
- provides easily quantifiable information

DISADVANTAGES

- does not provide the kind of information that would help *explain* success or failure
- fails to question goals themselves
- unintended outcomes may go unrecognized
- cannot be used for ongoing program improvement purposes

Decision-Oriented Evaluation

Assumes that evaluation is a function of program management, and its chief purpose is to inform decision-making and ensure accountability. Gathers information on any one or all of the following: what resources were invested in the program (inputs), what was actually done with those resources (processes), what were the results, both intended and unintended (outcomes), and what was the environment in which the program operated (context)? Employs a range of techniques, primarily (but not exclusively) quantitative.

ADVANTAGES

- generates questions, helps define the information needed for a decision
- can serve program improvement
- can look for unintended outcomes
- integrated into the management system
- can evaluate goals themselves
- typically provides easily quantifiable information

DISADVANTAGES

- information needs of managers drive the evaluation; other audiences may not be served
- can be costly and complex
- assumes that important decisions can be identified in advance and that the decision process is rational

(continued on reverse side)

Responsive or Naturalistic Evaluation

Assumes that the purpose of an evalution is to give all the people involved with a program the information they need to better understand and reflect on what they are doing. Centers less on goals and more on activities, transactions, and effects occurring in the program. Evaluation design and focus are typically not determined in advance; they emerge through observation and iteration. Employs "anthropological" techniques like direct observation, interviews, case studies, and document analysis to produce a rich, qualitative description.

ADVANTAGES

- problem- and issue-oriented
- illuminates unseen dimensions of program
- accommodates information needs of multiple audiences
- encourages reflection on program
- excellent vehicle for program improvement

DISADVANTAGES

- time-consuming and labor-intensive, therefore more costly and can be more disruptive
- results not generalizable to other programs in a traditional sense
- viewed by some as "soft"and "subjective" — may have less credibility

NOTE: The ideas and information presented here are excerpted from a recent report, *Testing and Evaluation: Learning from the Projects We Fund*, by George Madaus, Walt Haney, and Amelia Kreitzer. The full report is available for $15 prepaid, from the Council for Aid to Education, 51 Madison Avenue, Suite 2200, New York, New York 10010.

COUNCIL FOR AID TO EDUCATION

COUNCIL FOR AID TO
EDUCATION

51 Madison Avenue
Suite 2200
New York, New York 10010